

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Quaresma, M; (2020) Population-based cancer survival at small area level: methodological developments. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04658175>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4658175/>

DOI: <https://doi.org/10.17037/PUBS.04658175>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



**Population-based cancer survival
at small area level:
methodological developments**

Manuela Moreno Quaresma

Thesis submitted in accordance with the requirements
for the degree of Doctor of Philosophy
of the
University of London

2020

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

No funding received

Research group affiliation: Cancer Survival Group

Supervisors

Professor Bernard Rachet

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

Professor James Carpenter

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

Declaration of Authorship

I, Manuela Quaresma, declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

A black rectangular box redacting the signature of the author.

Date: 06/01/2020

This thesis is dedicated to

*the loving memory of my parents,
Maria Moreno and Manuel do Carmo*

*and to every person faced
with a cancer diagnosis*

Abstract

Cancer survival is a key indicator of the overall effectiveness of a health system in managing treatment and care of cancer patients. At the national level, cancer survival statistics facilitate overall surveillance of strategic importance. At the local level, they provide valuable insights into the performance of local cancer services essential for public health planning. There are however recurring concerns regarding both the estimation and dissemination of such survival outcomes, in particular at the smaller area level. The research presented in this thesis aimed to address some of these concerns.

A summary indicator of cancer survival, named Index of Cancer Survival, is proposed for all cancers combined designed to act as an overall measure of the effectiveness of cancer services in England at both national and local level. To estimate the index a two-step analytical approach was implemented, in which cancer survival is first estimated for each small area separately followed by a joint smoothing and mapping technique to filter out excessive variation from the resulting cancer survival maps. Such smoothed maps were thought suitable for national health policy-makers to devise national surveillance strategies, as they display in a simple way, the overall patterns of survival for the whole country. Funnel plots were then extended to visualise the spread of individual small-area cancer survival outcomes, mostly thought suitable for local health managers as a tool for monitoring the performance of survival outcomes in their local areas.

However, the estimation of cancer survival for small-health geographies remained challenging. The last part of this thesis explored how Bayesian approaches could be used to improve the estimation of cancer survival in the presence of sparse data, and when using more complex data structures, including spatially arranged and hierarchical data. The feasibility of an existing Bayesian model for the excess hazard using Poisson regression was explored to estimate small-area patterns in cancer survival accounting for the spatial structure of the data. A Bayesian flexible excess hazard regression model was then proposed based on the full likelihood specification to improve the modelling of both the baseline excess hazard and the smooth effect of continuous covariates using a special type of splines. The new model also accommodates hierarchical data allowing more complex cancer data structures to be modelled, such as patient level data nested within area of residence or hospital of care level data.

In summary, the cancer survival index and both data visualisation techniques for cancer survival greatly improved the interpretability and dissemination of such outcomes for non-technical audiences, in particular health policy-makers. Meanwhile, the Bayesian excess hazard model using Poisson regression improved the estimation when data were sparse by incorporating the spatial data structure. The Bayesian flexible excess hazard model in particular, enabled a better investigation of inequalities in cancer survival using a range of covariate effects and facilitated the study of more complex cancer data structures.

Acknowledgements

I would like to thank my main supervisor, Professor Bernard Rachet. I am very grateful for all his guidance, immense knowledge, support and patience throughout this very long process that is a part-time PhD. Mille mercis!

I am grateful to my supervisor Professor James Carpenter for his continuous encouragement and his invaluable input into the most complex technical aspects of my thesis. Thank you!

I am grateful to Professor Michel Coleman who received me with open arms in the Cancer Survival Group. His unparalleled knowledge about population-based cancer survival has taught me so much over the years. Obrigada!

To Mrs. Yuki Alencar, Cancer Survival Group Coordinator and a friend, thank you for all the help and support, and for being such a wonderful professional and person!

To all the colleagues of the Cancer Survival Group, past and present, and colleagues from LSHTM: Thank you so much to each and every one of you! I am grateful for all the words of encouragement, the many 'PhD chats' and the endless coffees!

To my friends and family for all their love and support. In particular to my sisters Francisca and Luisa for walking hand in hand with me through life as one soul, and to my parents Maria and Manuel for always encouraging me to study and look forward in life.

Preface

This thesis is submitted for the degree of Doctor of Philosophy at the London School of Hygiene & Tropical Medicine, University of London. The research presented herein was conducted under the supervision of Professor Bernard Rachet from the Department of Non-Communicable Disease Epidemiology and Professor James Carpenter from the Department of Medical Statistics, Faculty of Epidemiology and Population Health at the London School of Hygiene & Tropical Medicine.

I undertook my doctoral research as a part-time student, whilst simultaneously holding a full-time position as a Research Fellow in Statistics with the Cancer Survival Group, under the Cancer Research UK funded programme led by Professor Rachet. This programme investigates a vast range of research questions concerning socioeconomic inequalities in population-based cancer survival for patients diagnosed in England. During my studies, I also had the opportunity to provide analytical support for the production of official cancer survival statistics for the Office for National Statistics, and some ad hoc requests for cancer survival statistics from both national and local health policy-makers.

The research I present here, represent a series of interconnected questions, that arose from the research conducted within the programme and that needed to be addressed in a separate and more detailed manner. To address these research questions, I combined both my previous training as a medical statistician, with the research and practical experience I obtained from having worked with the Cancer Survival Group and the Southern Portuguese population-based cancer registry (prior to having joined the Cancer Survival Group), balancing the development of applied methodology for population-based cancer research with the implementation of tools for the dissemination of survival outcomes, particularly targeted at health policy-makers.

I hope that my doctoral research studies can contribute with a useful set of tools for cancer researchers across the world to implement national and local cancer survival monitoring tools, to analyse more complex cancer data, including spatial and hierarchical structures and to facilitate the dissemination survival outcomes to health policy-makers and other relevant non-expert audiences.

Manuela Quaresma
London, 2020

Contents

Declaration of Authorship	ii
Abstract	iv
Acknowledgements	v
Preface	vi
List of Figures	x
List of Tables	xii
Abbreviations	xiii
1 Introduction	1
1.1 Background	1
1.1.1 The burden of cancer	1
1.1.2 Cancer control programmes	2
1.1.3 Efficiency, equity and effectiveness of healthcare (E^3)	2
1.1.4 The NHS structure and English health geographies	3
1.1.5 Cancer survival deficit and inequalities in cancer survival	4
1.1.6 Demands for national and local monitoring of survival	5
1.2 Research aims and objectives	6
1.3 Research output: peer-reviewed publications	8
1.4 Thesis outline	8
2 Population-based cancer survival: overview of measures and estimators	9
2.1 What is population-based cancer survival?	9
2.2 'Classical' survival setting <i>versus</i> relative survival setting	10
2.3 Main measures of interest: Net survival and Excess hazard	11
2.4 Estimation of net survival and excess hazard	12
2.4.1 Non-parametric estimators	12
2.4.2 Regression models for the excess hazard	14

2.4.2.1	The likelihood function	15
2.4.2.2	Modelling the log-excess hazard	17
2.4.2.3	Modelling the cumulative log-excess hazard	17
2.4.2.4	Modelling the excess hazard using Poisson regression	19
2.4.3	Estimation of net survival using excess hazard regression models	20
3	Data for population-based cancer survival research	21
3.1	Population-based cancer registration	21
3.2	National Cancer Registration in England	22
3.2.1	Data items collected	23
3.2.2	Derived variables	24
3.2.2.1	Socioeconomic status	24
3.2.2.2	Health geographies	25
3.3	Additional data sources for cancer research	26
3.4	General population life-tables for cancer survival	26
4	An Index of Cancer Survival: a tool for national and local monitoring	27
4.1	Introduction	28
4.2	Specifications for the development of the index	28
4.3	Formulation of the index	29
4.4	Three-way standardisation technique	31
4.5	Standard weights for the estimation of the index	32
4.6	Application: Index of cancer survival for England and CCGs	43
4.6.1	Estimation of the individual index components	43
4.6.2	Developing a modelling strategy for the estimation of net survival	44
4.6.3	Model set-up	44
4.6.4	Selecting the number and location of spline knots	48
4.6.5	Model selection using the Akaike Information Criterion	49
4.6.6	Post-estimation of net survival	50
4.6.7	Dealing with model non-convergence	50
4.6.8	Combining the individual components of the index	51
4.6.9	Results 1: Index of cancer survival for England [1]	51
4.6.9.1	Research publication 1	56
4.6.10	Results 2: Index of cancer survival for CCGs [2–5]	72
4.7	Discussion	114
5	Data visualisation techniques for cancer survival relevant to health policy	117
5.1	Introduction	118
5.2	Smoothing technique for small-area cancer survival maps	119
5.3	Funnel plots for population-based cancer survival [6]	124
5.3.1	Research publication 2	125
5.4	Application: Smoothed maps and funnel plots to visualise the index of cancer survival for CCGs	139
5.4.1	Smoothed maps	139
5.4.2	Funnel plots	141
5.5	Discussion	143

6	Bayesian approaches for the estimation of cancer survival at small area level	145
6.1	Introduction	146
6.2	Overview of small-area estimation methods	148
6.3	Flexible Bayesian excess hazard models	154
6.4	Research publication 3	157
6.5	Research publication 4	182
6.6	Discussion	212
7	Discussion and Conclusions	214
A	Stata and R code	219
A.1	Stata code to estimate the national and local indexes of cancer survival . .	219
A.2	R code to construct a funnel plot	231
A.3	R code to implement flexible Bayesian excess hazard models using low-rank thin plate splines	234
B	Other relevant research activities undertaken	243
B.1	Research degree student poster day	243
B.2	Beautiful data competition	245
B.3	Oral presentations at conferences and meetings	247
B.3.1	North American Association of Central Cancer Registries	247
B.3.2	All-party Parliamentary Group on Cancer annual meeting	247
B.3.3	Royal statistical Society annual conference	247
C	Ethical approvals	261
	Bibliography	262

List of Figures

3.1	Map of the configuration of CCGs in England	25
4.1	Generic combinations needed for the estimation of the index of cancer survival using sex i ($i=1, 2$), age group j ($j=1, 2, \dots, J$) and cancer type k ($k=1, 2, \dots, K$).	30
4.2	Trends in the index of net survival for all cancers combined in England . . .	54
4.3	Net survival adjusted for age and sex for each cancer in 2010-11, and absolute change since 1971, England: 1, 5, and 10 years after diagnosis . .	54
4.4	Box-plots of one-year net survival index (%) for CCGs by calendar year of diagnosis: all adults, 1996-2011	73
5.1	Lung cancer incidence for women diagnosed in Finland (with permission to use from Prof. Pukkala at the Finnish cancer registry)	119
5.2	Map of lung cancer incidence for women diagnosed in Finland with overlaid raster grid	121
5.3	Circular window defining distance from center of grid point	121
5.4	Smoothing decay function	122
5.5	Funnel plot illustration	124
5.6	Smoothed maps of England using the one-year net survival index for CCGs	140
5.7	Funnel plots of the one-year net survival index for CCGs	142
6.1	Flow chart of data exclusions and hospital assignment after applying the algorithm to allocate the hospital of care or diagnosis.	190
6.2	Windrose graphs showing the distribution (%) of male patients diagnosed with colon cancer in London, 2006-2013: (a) least deprived versus most deprived category by CCG of residence; (b) least deprived versus most deprived category by hospital of cancer care; (c) stages at diagnosis 1, 2 and 3 versus stage 4 by CCG of residence (d) stages at diagnosis 1, 2 and 3 versus stage 4 by hospital of cancer care.	195
6.3	Flow map of London displaying the pathways of patients' journeys between the CCG of residence and the hospital of cancer care for men diagnosed with colon cancer, 2006-2013.	196
6.4	Flow map of London displaying the pathways of patients' journeys between the CCG of residence and the hospital of cancer care for women diagnosed with colon cancer, 2006-2013.	197

6.5	Funnel plots of 5-year net survival (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) by CCG of residence for stages at diagnosis 1, 2 and 3; (b) by hospital of cancer care for stages at diagnosis 1, 2 and 3; (c) by CCG of residence for stage at diagnosis 4; (d) by hospital of cancer care for stage at diagnosis 4.	198
6.6	Funnel plots of 5-year net survival (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) by CCG of residence for stages at diagnosis 1, 2 and 3; (b) by hospital of cancer care for stages at diagnosis 1, 2 and 3; (c) by CCG of residence for stage at diagnosis 4; (d) by hospital of cancer care for stage at diagnosis 4.	199
6.7	Funnel plots of 5-year net survival by CCG of residence (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stages at diagnosis 1, 2 and 3); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stages at diagnosis 1, 2 and 3).	206
6.8	Funnel plots of 5-year net survival by CCG of residence (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stage at diagnosis 4); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stage at diagnosis 4).	207
6.9	Funnel plots of 5-year net survival by CCG of residence (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stages at diagnosis 1, 2 and 3); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stages at diagnosis 1, 2 and 3).	208
6.10	Funnel plots of 5-year net survival by CCG of residence (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stage at diagnosis 4); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stage at diagnosis 4).	209
6.11	Funnel plots of the random effects by CCG of residence and hospital of care for women using: complete case analysis after removing cases with missing stage at diagnosis ((a) and (b)) and using all data by modelling the missing data structure ((c) and (d)).	210

List of Tables

4.1	First set of 'sex-age-cancer'-specific weights	34
4.2	Second set of 'sex-age-cancer'-specific weights	41
4.3	Location of knots for the spline on age at diagnosis for the England index	49
4.4	Location of knots for the spline on age at diagnosis for the CCG index	49
4.5	Models run for each cancer-sex combination (M1-M7, in the same order as specified in section 4.6.3); converged models marked with 'x'; and best model selected	52
4.6	England index: summary of best fitting models for all the cancer-sex combinations	53
4.7	CCG index: summary of best fitting models for all the cancer-sex combinations	72
4.8	Mid-year population estimates (thousands) for 2011, and number of cancer patients included in survival analyses, by calendar year of diagnosis 1996-2003: Clinical Commissioning Groups by regions, England	74
4.9	Mid-year population estimates (thousands) for 2011, and number of cancer patients included in survival analyses, by calendar year of diagnosis 2004-2011: Clinical Commissioning Groups by regions, England	84
4.10	One-year net survival index (NS: %) and precision of estimates (prec) for all cancers combined, by calendar year of diagnosis: all adults, Clinical Commissioning Groups, England, 1996-2003	94
4.11	One-year net survival index (NS: %) and precision of estimates (prec) for all cancers combined, by calendar year of diagnosis: all adults, Clinical Commissioning Groups, England, 2004-2011	104
6.1	Number of cases (N) and proportion of deaths (%) within the follow-up period by CCG of residence for men and women diagnosed with colon cancer in London, 2006-2013.	193
6.2	Number of cases (N) and proportions of deaths (%) within the follow period by hospital of cancer care for men and women diagnosed with colon cancer in London, 2006-2013.	194

Abbreviations

CSG	C ancer S urvival G roup
LSHTM	L ondon S chool of H ygiene & T ropical M edicine
ONS	O ffice for N ational S tatistics
WHO	W orld H ealth O rganisation
IARC	I nternational A gency for R esearch on C ancer
UICC	U nion for I nternational C ancer C ontrol
NHS	N ational H ealth S ervice
NCRAS	N ational C ancer R egistration and A nalysis S ervice
CCG	C linical C ommissioning G roup
PCT	P rimary C are T rust
HES	H ospital E pisode S tatistics
NBOCA	N ational B owel C cancer A udit
GLM	G eneralised L inear M odel
AIC	A kaïke I nformation C riterion
ML	M aximum L ikelihood
SAE	S mall A rea E stimation
CAR	C onditional A uto R egressive
SMR	S tandardised M ortality R ates
BYM	B esag Y ork and M ollié
GLMM	G eneralised L inear M ixed M odel
MCMC	M arkov C hain M onte C arlo
LRTP	L ow R ank T hin P late
APPGC	A ll- P arty P arliamentary G roup on C ancer

Chapter 1

Introduction

1.1 Background

1.1.1 The burden of cancer

The burden of cancer continues to increase worldwide, with an estimated 18 million new cancer cases diagnosed in 2018 and 9.5 million cancer deaths [7], along with marked variations in survival observed across all the regions of the globe [8]. In England, incidence rates are lower than in the European Union for men, but higher for women [7], with an estimated annual number of newly diagnosed cancers in 2016 of just over 300,000 cases [9]. Although prevention is preferable to cure, not all the cancers can be prevented. For a reduction in cancer mortality to occur, both a reduction in cancer incidence and an increase in cancer survival are essential [10]. Measuring the burden of cancer in the best possible way is a crucial aspect in cancer research, in which incidence, mortality and survival statistics (among others) are used to investigate the causes and outcomes of the disease [11–14]. The baseline information generated supports the development and improvement of cancer control strategies, as for instance, the prioritisation of regions for the implementation of cancer awareness [15] or early diagnosis campaigns to improve those health outcomes [16, 17]. Understanding how cancer impacts a population and varies between different populations, and over time is thus essential for public health planning and surveillance [18, 19].

1.1.2 Cancer control programmes

Cancer control programmes, cancer plans or cancer strategies as defined by the World Health Organisation (WHO) are public health programmes designed to reduce the number of cancer cases and deaths in a population, and to improve the prognosis and the quality of life of cancer patients. They are based on the implementation of systematic, equitable and evidence-based strategies for prevention, early detection, diagnosis, treatment and palliation of all cancer cases in a population [20].

In 2013, the Union for International Cancer Control (UICC) launched the World Cancer Declaration, calling upon all government leaders and health policy-makers to reduce the global cancer burden, promote greater equity, and integrate cancer control into the world health and development agenda. The declaration set out one overarching goal: *'There will be major reductions in premature deaths from cancer, and improvements in the quality of life and cancer survival'* [21].

1.1.3 Efficiency, equity and effectiveness of healthcare (E^3)

Efficiency, equity and effectiveness, the so-called 3 E 's, are three terms used in healthcare performance evaluation [22]. Efficiency can be defined as the allocation of the limited economic resources to meet the healthcare needs of the population at minimum costs, although there are many other ways in which efficiency can be defined, as for instance, the maximisation of health benefits from available resources [23]. Equity refers to the fair distribution or allocation of the resources within the healthcare system [24, 25]. Balancing efficiency and equity has been the biggest dilemma for the National Health Service (NHS) in England since its origin [26], in particular for those policies that increase efficiency but also increase health inequalities, or improve equity whilst decreasing efficiency.

The NHS is the publicly funded national healthcare system in England. It was created in 1948, based on the core principle of equity, setting out that good healthcare should be available to all, regardless of their wealth [27]. The three core principles that have guided the implementation and development of the NHS over the last 70 years remain the same: 1) that it meets the needs of everyone; 2) that it be free at the point of delivery; and 3)

that it be based on clinical need, not ability to pay. Equity within the NHS has been guided mainly by an equal access principle rather than equal outcome. In an effort to reduce the persistent and widening health inequalities that have been reported in England since the 1980s [28], equity has been put at the centre of all subsequent health policies.

Effectiveness relates to the extent to which certain policies on healthcare provision achieve their intended purposes, in terms of best and equal outcomes for all the patients. It can be measured in terms of resource allocation or access to healthcare, but an alternative way to measure effectiveness is to use an outcome-based measure such as a cancer survival. Quantifying disparities in cancer survival will enable the identification of differences in the effectiveness of cancer patient care within the healthcare system [29].

1.1.4 The NHS structure and English health geographies

The NHS is formed by many independent bodies and sub-organisations as laid out by the Health and Social Care Act [27, 30, 31]. The 2013 NHS restructuring created organisations such as the *NHS England* and the 211 geographically-based *Clinical Commissioning Groups* (CCGs). NHS England is responsible for commissioning the planning and buying of healthcare services, such as primary care services, and setting the priorities and direction of the NHS. It also allocates 60% of the NHS budget to CCGs across England. CCGs are clinically led statutory NHS bodies responsible for the planning and commissioning of healthcare services, including General Practitioner (GP) services, planned hospital, urgent and emergency care.

Prior to the creation of CCGs, several other health geographies were central in the organisation of the NHS. Primary Care Groups have been created in 1999 to act as units of local organisation for healthcare delivery in England. The initial 481 groups, were restructured in 2001 to become 303 Primary Care Trusts (PCTs). These were further reduced to 152 PCTs in 2006 and to 151 PCTs two years later [32], before being abolished and replaced by CCGs in 2013. Unlike PCTs, who held the budgetary responsibility for the delivery of care for patients living in their catchment area, CCGs only hold the responsibility for managing the healthcare delivery of patients registered in their practices.

The numerous changes to English health geographies over the last decades [33], through mergers, boundary changes, creation and cessation of geographies have complicated the production of official cancer survival statistics [34]. These changes have in particular prevented the availability of long-term survival trends for the affected geographies that could provide valuable insights into the survival improvements in those areas.

1.1.5 Cancer survival deficit and inequalities in cancer survival

Since the mid-1990s, the level of cancer survival in England has been documented to consistently fall below the European average, lower than most Western European countries [35–38], and some non-European countries considered to be equivalent in terms of wealth and healthcare organisation [39]. Several studies have also shown wide geographical variations within England for most of the common adults cancer types, including a persistent North-South gradient, with lower survival in the North of England, suggesting that the place of residence plays an important role in the survival of a cancer patient [18, 40–48]. In addition, wide inequalities in cancer survival by socio-economic status have been extensively described in England [46, 49–54], despite the existence of universal access to care within the equity based NHS.

The widely documented English cancer survival deficit has generated much debate within the political and health communities in the last three decades [55–57], to the extent that improving quality of care, setting targets, increase investment and improving survival became a top priority for the Government, working together with cancer charities and research groups to achieve these objectives [57]. Since the Calman-Hine report was released in 1995, launching a policy framework for commissioning cancer services [58] and with the subsequent implementation of the NHS Cancer Plan in 2000, a series of other national initiatives have since reinforced the need for an integrated strategy to tackle cancer inequalities and improve prevention, early diagnosis and survival in England [59–61]. As summarised by Alan Milburn, Secretary of State for Health at the time of the introduction of the NHS cancer plan [62, 63]:

'The poor are still far more likely to get cancer than the rich, and their chances of survival are lower too. Furthermore there are too many variations in the quality of care and treatment across the country, leaving cancer patients frustrated by a postcode lottery.'

The cancer reform strategy, first introduced in 2007, [60, 64–66] stated along with other points that:

1. *'Cancer networks will support Primary Care Trusts in commissioning high quality, safe and effective cancer services.'*
2. *'Tools will be made available to Primary Care Trusts to enable them to commission effectively and benchmark their performance...'*

However, at the start of the research presented in this thesis, no official tools had been defined to monitor and assess the cancer plan, and the reform strategy in terms of survival outcomes for the most recent health geographies configurations.

1.1.6 Demands for national and local monitoring of survival

More recent initiatives, such as the NHS Long Term Plan for cancer [67] published in early 2019, building on the Independent Cancer Taskforce strategy [68] published in 2015, reinforce their ambitions and commitments to improve cancer outcomes in England over the next ten years, in particular aiming at increasing the number of people surviving their cancer for at least five years after diagnosis. As a result, monitoring improvements in these outcomes in a timely and systematic way became even more crucial, leading to increased demands and pressures from national policy-makers and local healthcare managers to have available cancer survival monitoring tools for the most recent configurations of health geographies [69]. National health policy-makers are mainly interested in understanding the overall patterns of survival to help devise national surveillance strategies, whilst local healthcare managers, are more interested in understanding how their local catchment area, i.e. the health geography of their responsibility is performing and how they compare with the national average performance. Addressing such demands in the best possible way depends on several factors: a) the availability of adequate statistical methods for the estimation of

cancer survival, in particular of methods that enable the analysis of more challenging data structures, including for instance, smaller health geographies; b) the timely availability of individual-level cancer patient data and other relevant healthcare system level information; and c) the existence of visualisation tools for a more effective dissemination of survival outcomes to non-expert audiences.

1.2 Research aims and objectives

The overarching aim of my doctoral research studies is to provide tools enabling robust estimation of cancer survival and effective communication of survival outcomes. This thesis is further divided into three main aims and specific objectives as described below.

Research aim 1

To summarise and monitor survival for all cancers combined in England at both national and local level.

The specific objectives to achieve this aim are:

- 1.1 To design a summary survival indicator for all cancers combined using a three-way standardisation technique;
- 1.2 To create suitable sets of weights to estimate the summary indicator;
- 1.3 To implement a modelling strategy to estimate the individual cancer survival components needed for the summary indicator using excess hazard regression models;
- 1.4 Application 1: to estimate 40-year trends in the summary survival indicator using patients diagnosed with cancer between 1971-2011 in England, i.e. estimate the summary indicator at national level;
- 1.5 Application 2: to estimate 16-year trends in the summary survival indicator using patients diagnosed with cancer between 1996-2011 in each of the 211 CCGs in England, i.e. estimate the summary indicator at local level.

Research aim 2

To improve the visualisation of cancer survival for a more successful dissemination to policy-makers.

The specific objectives to achieve this aim are:

- 2.1 To adapt a joint smoothing and mapping technique for cancer survival that produces smooth map surfaces based on small-area survival estimates;
- 2.2 To extend the use of funnel plots to visualise the spread of individual cancer survival estimates around a pre-specified target value by formulating the correct control limits for cancer survival;
- 2.3 Application: to use these two techniques to visualise the results of the index of cancer survival by CCG (estimated in Aim 1), and to exemplify how the same set of results can be used for national surveillance and for local monitoring of cancer survival.

Research aim 3

To determine how Bayesian approaches can be used in the cancer survival setting to improve the estimation of survival in the presence of sparse data, and when using more complex data structures, including spatially arranged and hierarchical data.

The specific objectives to achieve this aim are:

- 3.1 To summarise the existing literature for the estimation of cancer survival in the presence of sparse data;
- 3.2 To propose a flexible Bayesian excess hazard model formulated on the log-excess hazard scale, and to demonstrate how net survival can be estimated from such a model;
- 3.3 Application: to extend the model proposed in objective 3.2 to accommodate random effects and use this model to investigate variation in net survival for patients diagnosed with colon cancer living and receiving care in London.

1.3 Research output: peer-reviewed publications

Four publications have been prepared based on the research conducted during my doctoral studies. Three have been published and the last one is ready to be submitted for publication.

Research publication 1 Quaresma M, Coleman MP and Rachet B. 2015. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *The Lancet*, 385, 1206-1218.

Research publication 2 Quaresma M, Coleman MP and Rachet B. 2014. Funnel plots for population-based cancer survival: principles, methods and applications. *Statistics in Medicine*, 33, 1070-1080.

Research publication 3 Quaresma M, Carpenter J and Rachet B. 2019. Flexible Bayesian excess hazard models using low-rank thin plate splines. *Statistical Methods in Medical Research*. Published online first September 2019.

Research publication 4 Quaresma M, Carpenter J, Turculet A and Rachet B. Variation in survival for patients diagnosed with colon cancer living and receiving care in London, 2006-2013: does it matter where you live? - ready to be submitted to *The Lancet*.

1.4 Thesis outline

The remainder of this thesis is organised as follows. Chapter 2 introduces the main concepts for population-based cancer survival. Chapter 3 describes the main sources of data available for cancer survival research. The main research chapters for aims 1, 2 and 3 are presented in Chapters 4, 5 and 6, respectively. In each chapter the research publications are intertwined with additional unpublished material relevant for the completeness of the research presented. Chapter 7 makes some concluding remarks and suggests future lines of research. The appendices present Stata and R code for all the analysis performed, describe other relevant activities undertaken and give information regarding the ethical approvals obtained for this research.

Chapter 2

Population-based cancer survival: overview of measures and estimators

This chapter presents an overview of the main measures and estimators for population-based cancer survival research. Specific details of the methods used throughout this thesis are given within each of the relevant research chapters.

2.1 What is population-based cancer survival?

Population-based cancer survival methodology refers to a collection of methods developed to study the time between the diagnosis of cancer and the death of a patient [70]. The event of interest is defined as the death due to cancer and the main aim is to quantify the survival of a cohort of cancer patients that can be attributed only to the cancer of interest. Data for population-based cancer survival research are routinely collected by population-based cancer registries, which record information on all the cancer cases diagnosed within their area of catchment, and thus cancer survival measured from such data represents the whole population. Contrasting with survival estimates obtained from clinical trials, which are

designed to quantify the highest achievable survival, using data from population-based cancer registries will quantify the average survival achieved in that population.

Cancer survival is interpreted as the survival patients would experience if the cancer of interest was the only possible cause of death. Some authors relate this interpretation to a *hypothetical world* where patients cannot die from anything else. In the *real world*, patients die from any cause, including their cancer, and thus such a measure is not directly relatable to individual cancer patients since it does not quantify their prospects of survival. Cancer survival is a measure of cancer prognosis, tailored for health policy to support the evaluation of the effectiveness of healthcare systems in managing cancer. Cancer survival can be monitored over time within the same population or compared between populations, without having the influence of other causes of death, in particular without the influence of unequal distributions of other causes between populations.

2.2 ‘Classical’ survival setting versus relative survival setting

‘Classical’ survival methodology provides a suit of methods for the estimation of time-to-event (survival) outcomes [71, 72]. Overall survival for a cohort of cancer patients is estimated defining as events all the deaths, regardless if cancer related or not. Such a measure is not relevant in population-based cancer survival research for the reasons mentioned in the previous section. Survival can also be derived from the cause-specific mortality hazard where a specific cause (here the cancer) is the only event of interest, whereas the events due to other causes are censored in order to remove such competing risks. This estimator is of limited use to estimate population-based cancer survival due to two main reasons [73]. First, to define the cause of death due to cancer as the event of interest, the cause of death needs to be known for each cancer patient. This is rarely the case at population-level, since the cause of death is often unknown or, in the absence of a single study protocol (such as in the clinical trial context), unreliable [74, 75]. Second, an important assumption in survival analysis is that the censoring process is assumed to be independent from the process that generates the events, i.e. the censoring is non-informative [71]. The process becomes informative when one or more factors influence both

mortality hazards: the cancer-specific hazard and the other-causes hazard, leading to biased estimates of survival [76]. For example, older patients are more likely to be censored than younger patients due to other causes of death, making the censoring process informative and leading to biased estimates of cancer survival [77, 78].

The framework (or setting) of *relative survival* was introduced in the 1950s to address these challenges, defining a set of measures for population-based cancer survival. Many estimators have since been proposed within this framework, all of which do not require the information about the cause of death to be known [79, 80]. In contrast with the ‘classical’ survival setting, in which only observed event times, event indicators and in some instances, patient- and tumour-level characteristics are used in the estimation of survival, in the relative survival setting the all-cause mortality rates from the general population are also used as an estimate of the competing risks of deaths due to causes other than cancer.

2.3 Main measures of interest: Net survival and Excess hazard

The two main measures of interest to be quantified in population-based cancer survival are called net survival and the excess hazard. Net survival is defined as the survival that would occur in a cohort of cancer patients if the cancer of interest was the only cause of death that patients could experience [70, 81]. The excess hazard is defined as the hazard of death that can be attributed only to the cancer of interest, i.e. in ‘excess to’, or after accounting for all the other causes of death present in the general population from where the cancer patients originated.

Similarly to the classical survival setting, where several useful relationships can be established between the survival function and the hazard function, relationships hold between the net survival function and the excess hazard function. Generically, T being a non-negative random variable representing the observed survival times t , we define

$$H_E(t) = \int_0^t h_E(u) du \quad (2.1)$$

$$S_{net}(t) = \exp(-H_E(t)) \quad (2.2)$$

$$H_E(t) = -\log(S_{net}(t)) \quad (2.3)$$

where $H_E(t)$ is the cumulative excess hazard function at time t , $h_E(t)$ is the excess hazard function at time t and $S_{net}(t)$ is the cumulative net survival function at time t .

2.4 Estimation of net survival and excess hazard

Several estimators have been proposed for net survival and for the excess hazard within the relative survival setting. In the next sections, a summary of the main estimators for these two measures is presented.

2.4.1 Non-parametric estimators

Three non-parametric estimators, *Ederer I* [82], *Ederer II* [83] and *Hakulinen* [84], were proposed for net survival between the 1960s and the early 1980s. These estimators follow the same generic formulation, so-called ‘ratio-estimators’

$$S_{net}(t) = \frac{S_O(t)}{S_P(t)} \quad (2.4)$$

where $S_{net}(t)$ is the net survival at time t , $S_O(t)$ is the observed survival for the cancer patient cohort at time t , and $S_P(t)$ is the expected survival or background population survival for the cancer patient cohort at time t , i.e. the survival of the cohort if patients were disease-free.

The observed survival in the numerator, $S_O(t)$, is estimated using classical survival estimators, i.e. Kaplan-Meier or actuarial methods [85–87] defining as event of interest all the deaths in the cancer patient cohort, regardless if cancer related or not. The expected survival in the denominator, $S_P(t)$, is calculated using all-cause mortality rates from the

general population. The three estimators differ in the way the expected survival is calculated, in successive attempts to best estimate net survival and to satisfy the non-informative censoring assumption. Throughout the decades the term ‘relative survival estimates’ was used in all the literature, although the original aim of these three estimators was to quantify net survival.

In 2012, Pohar-Perme et al. [77] laid out these three estimators examining their properties and defining what quantity they were estimating. The authors concluded that none of the three estimators was adequately estimating net survival and that the non-informative assumption was not taken into account correctly. The authors proposed a new estimator that takes into account the informative censoring using as weights the probability of each patient remaining at risk of death in the general population. Subsequent work compared the performance of the new ‘Pohar-Perme’ estimator with the previous three estimators confirming the magnitude of the biases introduced by the latter ones [88]. The ‘Pohar-Perme’ estimator became the current gold standard for non-parametric estimation of net survival [89].

‘Pohar-Perme’ net survival estimates can be calculated for the whole cohort or stratified by different levels of categorical variables using the R command *relsurv* [90] or the Stata command *stns* [91]. Both interval-specific and cumulative probabilities of net survival can be estimated for specific times or for the entire follow-up time defined. Non-parametric estimates of the excess hazard can also be derived using the ‘Pohar-Perme’ estimator, but the excess hazard is more commonly estimated using regression models as described in the next section.

2.4.2 Regression models for the excess hazard

Modelling time-to-event data offers the possibility of investigating the effect of multiple prognostic factors (or covariates) on the form of the hazard function, as well as estimating the hazard for an individual patient or group of patients [72]. The use of regression models in the ‘classical survival’ setting became widespread in medical research with the introduction of the popular Cox proportional hazards model [92], which in its semi-parametric form quantifies hazard ratios without having to specify a parametric form for the baseline hazard. Other regression models have since been proposed in the literature for survival data, including fully parametric models, such as the widely used Weibull regression model and other more flexible alternatives using high dimensional polynomials, as for example spline functions or fractional polynomials [93–98].

Regression models for the excess hazard defined within the relative survival setting are based on the additive decomposition of the total (or overall) hazard into two components: the hazard due to the cancer of interest (the excess hazard) and the hazard due to all other causes of death in the general population (the expected population hazard or background mortality),

$$h(t) = h_E(t) + h_P(t) \quad (2.5)$$

where $h(t)$ is the total or overall hazard at time t , $h_E(t)$ is the excess hazard due to the cancer at time t and $h_P(t)$ is the general population hazard at time t .

Estève et al. [81] introduced the first regression model for the excess hazard based on the full-likelihood specification using individual survival time data. In its original formulation, this model was proposed on the log-excess hazard scale with the baseline log-excess hazard modelled as a piecewise constant (or step function). Several extensions and refinements have been proposed both to the model on the log excess hazard scale [99], as well as introducing models defined on the log cumulative excess hazard scale [100] and formulations based on Generalised Linear Models [101]. The proposed models mainly allowed the non-proportionality and non-linearity assumptions to be relaxed for covariates and interaction terms, and the baseline excess hazard to be modelled with flexible functions to avoid the clinically implausible jumps in the hazard imposed by a piecewise constant function. The

most common flexible functions used in excess hazard modelling are splines, in particular B-splines [99] and restricted cubic splines [100], although fractional polynomials have also been proposed [102].

The next section will introduce three of the most common formulations for excess hazard models: 1) the model formulated on the log-excess hazard scale; 2) the model formulated on the cumulative log-excess hazard scale; and 3) a Generalised Linear Model (GLM) formulation modelling the number of observed deaths using Poisson regression. These three models will be used in the research presented throughout the thesis, and the choice of models is related to the availability of ready-to-use software, both in Stata [103] and R software [104], at the time the different analysis were performed. Before introducing each of the models, we start by formulating the likelihood function for a generic excess hazard model.

2.4.2.1 The likelihood function

Let $(t_i, \mathbf{x}_i, \delta_i)$, $i=1, \dots, n$, $t_i > 0$, denote a set of n time to event observations, measured from the date of diagnosis of a cancer until death, with covariates \mathbf{x}_i and vital status indicator δ_i ($\delta_i=0$ if censored, $\delta_i=1$ if death occurred). The likelihood function of the full vector of parameters of interest $\theta = (\theta_1, \theta_2, \theta_3, \dots)$ is written in generic terms as

$$L(\theta) = \prod_{i=1}^n h(t_i, \mathbf{x}_i, \theta)^{\delta_i} \cdot S(t_i, \mathbf{x}_i, \theta) \quad (2.6)$$

where $h(t_i, \mathbf{x}_i, \theta)$ is the hazard function and $S(t_i, \mathbf{x}_i, \theta)$ is the survivor function for an observation t_i .

Considering only the individual contribution of observation t_i to the log-likelihood, the function can be written as

$$\log L(\theta) = \delta_i \cdot \log(h(t_i, \mathbf{x}_i, \theta)) + \log(S(t_i, \mathbf{x}_i, \theta)) \quad (2.7)$$

Using the following relationship between the survival function and the cumulative hazard function as expressed in equation (2.3)

$$\log(S(t_i, \mathbf{x}_i), \theta) = -H(t_i, \mathbf{x}_i, \theta) = - \int_0^{t_i} h(u, \mathbf{x}_i, \theta) du \quad (2.8)$$

and replacing equation (2.8) into equation (2.7), the contribution of observation t_i to the log-likelihood can be rearranged as

$$\log L(\theta) = \delta_i \cdot \log(h(t_i, \mathbf{x}_i), \theta) - \int_0^{t_i} h(u, \mathbf{x}_i, \theta) du \quad (2.9)$$

Considering the additive decomposition of the total hazard, $h(t_i, \mathbf{x}_i, \theta)$, into the two components as defined in equation (2.5)

$$h(t_i, \mathbf{x}_i, \theta) = h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i) \quad (2.10)$$

where $h_E(t_i, \mathbf{x}_i, \theta)$ is the excess hazard function for an observation t_i and $h_P(a_i + t_i, \mathbf{z}_i)$ is the population hazard function for an observation t_i , evaluated at the attained age at death (or age at censoring): $a_i + t_i$, with a_i the age at diagnosis and \mathbf{z}_i ($\mathbf{z}_i \subseteq \mathbf{x}_i$) a subvector of covariates. The population hazard is assumed to be a known quantity, taken as the age-specific mortality rates from population life tables, stratified as finely as possible according to \mathbf{z}_i , possibly including, in addition to age at death (or censoring), gender and calendar year, socio-economic status, ethnicity or region of residence [105].

Replacing equation (2.10) into equation (2.9), the log-likelihood can be rewritten entirely (up to a constant) as a function of the excess hazard and the population hazard

$$\log L(\theta) \propto \delta_i \cdot \log[h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i)] - \int_0^{t_i} h_E(u, \mathbf{x}_i, \theta) du \quad (2.11)$$

Given that the population hazard $h_P(a_i + t_i, \mathbf{z}_i)$ is assumed to be a constant, inferences based on equation (2.11) are made by specifying an appropriate model for the excess hazard function $h_E(t_i, \mathbf{x}_i, \theta)$.

2.4.2.2 Modelling the log-excess hazard

Estève et al. [81] proposed the first regression model on the log-excess hazard scale modelling the baseline log-excess hazard as a piecewise constant step function. Several flexible functions have since been proposed to extend this model [99, 106–109], which assumes a multiplicative effect of the covariates on the baseline excess hazard ($h_{E_0}(t)$) as

$$h_E(t, \mathbf{x}, \theta) = h_{E_0}(t) \cdot \exp(\theta \cdot \mathbf{x}) \quad (2.12)$$

The model for the logarithm of the excess hazard can be written in generic terms by taking the logarithm of equation (2.12) as

$$\log(h_E(t, \mathbf{x}, \theta)) = \log(h_{E_0}(t)) + \theta_1 \cdot x_1 + g_1(x_2) + g_2(t) \cdot x_3 + \dots \quad (2.13)$$

where $\log(h_{E_0}(t))$ is now the baseline log-excess hazard function; θ_1 is a linear and proportional effect on the log-excess hazard of covariate x_1 ; $g_1(x_2)$ is a non-linear and proportional effect of a continuous covariate x_2 ; $g_2(t)$ is a non-proportional (i.e. time-dependent) effect of a covariate x_3 .

Options available to fit models on the log-excess hazard scale include the Stata command *strel* (for the original Estève et al. model) [110], the Stata command *strcs* [111] and the R command *mexhaz* [112].

2.4.2.3 Modelling the cumulative log-excess hazard

Flexible parametric regression models have been proposed within the ‘classical’ survival setting by Royston and Parmar [96]. These models use restricted cubic splines to model the baseline cumulative hazard and the smooth effect of covariates and interaction terms [113]. Restricted cubic splines are piecewise polynomials that join at points called internal knots. They are restricted to be linear before the first knot and after the last knot to improve fit in the tails of the distribution. The choice of the location of the internal knots is not always intuitive, and a default location is often suggested based on the centiles of the event distribution, without compromising the goodness of fit of the models [100].

The model set-up is derived from a common parametrisation of the Weibull survival function as follows. Considering

$$S(t) = \exp(-\lambda t^\gamma) \quad (2.14)$$

with $S(t)$ the survivor function for a Weibull distribution at time t and λ and γ the scale and shape parameters of the distribution, respectively.

Transforming equation (2.14) into the log cumulative hazard scale

$$\ln(H(t)) = \ln(-\ln(S(t))) = \ln(\lambda) + \gamma \ln(t) \quad (2.15)$$

And adding a set of covariates ($\theta \cdot \mathbf{x}$) to the formulation, rearranges into

$$\ln(H(t, \mathbf{x}, \theta)) = \ln(\lambda) + \gamma \ln(t) + \theta \cdot \mathbf{x} \quad (2.16)$$

The Royston and Parmar model originates from this formulation by relaxing the assumption that the log cumulative baseline hazard function ($\ln(\lambda) + \gamma \ln(t)$) is linear on the log-time scale, and by modelling it with a spline adding more flexibility to the model. The model is re-written as

$$\ln(H(t, \mathbf{x}, \theta)) = s(\ln(t)|\gamma, kn) + \theta \cdot \mathbf{x} \quad (2.17)$$

where $s(\ln(t)|\gamma, kn)$ is a restricted cubic spline of $\ln(t)$ and kn are the spline knots.

Nelson et al. [114] have extended the Royston and Parmar model to the relative survival setting, by considering the decomposition of the total cumulative hazard as

$$H_T(t) = H_{net}(t) + H_{exp}(t) \quad (2.18)$$

where $H_T(t)$ is the cumulative total hazard at time t , $H_{net}(t)$ is the cumulative excess hazard at time t and $H_{exp}(t)$ is the cumulative expected hazard.

The log cumulative excess hazard is model is formulated as

$$\ln(H_{net}(t, \mathbf{x}, \theta)) = s_0(\ln(t)|\gamma, kn) + \theta \cdot \mathbf{x} \quad (2.19)$$

where $s_0(\ln(t)|\gamma, kn)$ is now the restricted cubic spline modelling the baseline cumulative log-excess hazard function.

This model can be fitted using the Stata command *stpm2* [100].

2.4.2.4 Modelling the excess hazard using Poisson regression

An alternative regression model for the excess hazard was proposed by Dickman et al. [101] formulated as a generalised linear model (GLM) with a Poisson error structure for the observed number of deaths [115] and a piecewise constant excess hazard. A lexis expansion of the data is required at pre-specified cut-off points of the follow-up period into i intervals. The number of deaths d_i in each follow-up interval i is modelled defining

$$d_i \sim \text{Poisson}(\mu_i) \text{ and } \mu_i = \lambda_i y_i \quad (2.20)$$

with intensity μ_i defined by λ_i the excess hazard function for the i^{th} interval and y_i the person-time at risk.

A dedicated link function proposed by Dickman et al. for the excess hazard formulation [101] is incorporated as

$$\log(\mu_i - d_i^*) = \log(y_i) + \theta \cdot \mathbf{x} \quad (2.21)$$

where $\log(y_i)$ is the offset and d_i^* is the expected number of deaths in interval i .

This model can be fitted by first using the Stata commands *stsplit* and *strs* [116] to perform the lexis expansion and to calculate the number of deaths in each interval i , the expected number of deaths and the person-time at risk, followed by the conventional *glm* command.

2.4.3 Estimation of net survival using excess hazard regression models

Excess hazard regression models as described in the previous sections can be used to derive model-based estimates of net survival for the whole cohort. Danieli et al. [117] have recommended that variables which can compromise the non-informative censoring assumption should be included in the model formulation when the aim is to estimate net survival. The main variables are those commonly defined in the cancer patient population and in the population life-tables, such as age, socioeconomic status and other relevant variables by which the life-tables are stratified.

After fitting an excess hazard model, post-estimation procedures are used to predict 'individual' excess hazard functions for each observation. From these, and using the combined equations (2.1) and (2.2), an 'individual' net survival function is derived for each observation as

$$S_{net_i}(t) = \exp\left(-\int_0^t h_{E_i}(u)du\right) \quad (2.22)$$

where $S_{net_i}(t)$ is the net survival function for observation i and $h_{E_i}(t)$ is the corresponding excess hazard function for observation i .

Net survival for the whole cohort, i.e. the marginal net survival function, is derived by averaging the 'individual' net survival functions for all the observations in the cohort as

$$S_{net}(t) = \frac{1}{n} \sum_{i=1}^n S_{net_i}(t) \quad (2.23)$$

where $S_{net}(t)$ is the marginal net survival function for the whole cohort.

The same procedure can be used to derive marginal net survival for subgroups of the cohort by averaging the 'individual' net survival functions within those subgroups.

Chapter 3

Data for population-based cancer survival research

This chapter introduces the main sources of data available for cancer research and defines the main variables of interest. Specific details regarding the extracts of data used throughout this thesis are described in each of the relevant chapters.

3.1 Population-based cancer registration

Population-based cancer registration is an essential component of any cancer control programme. The main function and responsibility of a cancer registry is to systematically collect and classify information on all the cancers that occur within a well-defined population, also known as catchment area [118]. This procedure is well established in many countries throughout the world and follows guidelines set by entities such as the International Agency for Research on Cancer (IARC), UICC and WHO to ensure the best quality and completeness of information [119, 120]. The data collected is used in a wide range of cancer control areas, from aetiological studies, primary and secondary prevention, to planning, monitoring and evaluation of cancer services and outcomes [121].

3.2 National Cancer Registration in England

In England, a regional cancer registration system covering the whole country started collecting data in 1962. The National Cancer Registry, then based at the Office for National Statistics (ONS) collated the regional datasets to compile a national cancer data repository and performed strict data quality checks before the data were released for analysis. Since 2016, the National Cancer Registration system was transferred from the ONS to Public Health England, that now maintains the National Cancer Registration and Analysis Service (NCRAS) to the same level of excellence regarding data quality procedures. Annual reports on cancer registration statistics, include performance indicators on data completeness, proportion of Death Certificate Only (DCO) registrations, records with zero survival, and proportion of cancers that are microscopically verified to ascertain their malignancy.

Cancer registration is a dynamic process. The data repository is regularly updated and revised, by either adding missed cancer registrations, deleting errors or updating information on existing cancer records as it becomes available. Since each patient can be diagnosed with more than one cancer, the data repository is based on individual cancer records instead of cancer patients. A multiple cancer identifier connects each patient to their multiple cancer records. A new cancer registration can only be completed after a minimum of six months following the date of diagnosis to allow treatment information to become available. In England, the target to complete cancer registration for a given calendar year is within 12 months from the end of that year. DCO cases only have a date of death recorded and miss a date of diagnosis to become a full cancer registration. These cases are excluded from further analysis since a survival time cannot be calculated. Some DCO cases can successfully be traced back to medical notes to retrieve a date of diagnosis, becoming a full cancer registration flagged as Death Certificate Initiated (DCI). Such records are included in further analysis. Cancer records for which the diagnosis occurred on the same day of death are flagged as 'true zero survival'. For these records a small amount of time, usually a day, is added to their survival time in order to include them in survival analysis [122].

3.2.1 Data items collected

The National cancer registration service collects a minimum set of information for every cancer registration, including patient demographics, tumour characteristics and type of treatment. The data items collected include: postcode of residence, gender, date of birth, date of cancer diagnosis, date of death, anatomic location (cancer site codes based on the International Classification of Diseases (ICD) [123]), morphological type (morphology codes based on the International Classification of Diseases for Oncology (ICD-O) [124]), behaviour of tumour, multiple tumour indicator, site (laterality), tumour grade, death certificate only indicator, treatment indicators (for surgery, radiotherapy, chemotherapy, hormonal therapy and others) and stage of disease at diagnosis for selected cancers. This minimum set of information collected for the whole cancer patient population aims to maximise record completeness and greater accuracy. Information is obtained from several sources, including hospital records, pathological reports, cancer treatment departments and General Practitioners medical files.

Since 1971, the National Cancer Registry dataset is routinely linked to the National Health Service Central Register (NHSCR), that updates every individual cancer record with information about the vital status of patients, flagging them as alive, emigrated, dead or not traced. During the 1970s and 1980s, over 96% of all registered cancer patients were successfully traced through the NHSCR, reaching over 99.6% since the 1990s. This follow-up procedure by electronic linkage is defined as passive follow-up. This contrasts with active follow-up procedures in operation in other countries, where patients are followed-up by direct contact, including phone calls and home visits. In settings where the recording of deaths is statutory and data linkage with cancer data is authorised, passive follow-up is the preferred follow-up method in cancer registration.

The Cancer Survival Group at the London School of Hygiene and Tropical Medicine holds the complete cancer registry database for individual patients diagnosed in England between 1971-2014 and followed up until 2015. This database was available for the research presented in this thesis. Specific data extracts used will be described in each of the relevant chapters.

3.2.2 Derived variables

Additional variables are derived for each cancer record based on the information collected in the cancer data repository. The two most relevant variables derived for this research are socioeconomic status and health geographies as detailed below.

3.2.2.1 Socioeconomic status

The English cancer registry does not collect information regarding the individual socioeconomic status (or deprivation) of cancer patients. Instead, an ecological measure of deprivation is allocated to each patient based on their postcode of residence at the time of their diagnosis. Several deprivation indexes have been developed based on census or administrative data, including the Carstairs Deprivation Index, the Townsend Deprivation Index and the Index of Multiple deprivation (IMD) [125–129]. Each cancer patient's record is linked to the smallest possible geography at which each deprivation index is defined. Taking the income domain scores of the IMD as an example, patients are categorised into five groups, from least deprived (category 1) to most deprived (category 5), according to the quintiles of the national distribution defined at the Lower Super Output Area (LSOA) level (~ 1,500 inhabitants). It has been shown that the choice of deprivation index has little impact on differences in cancer survival by deprivation, and that it is more important that the underlying geographies for which the indexes are defined are as small as possible to provide a good proxy of the individual deprivation of patients [130, 131]. The Townsend Deprivation Index is used for patients diagnosed between 1971 and 1986, the Carstairs Deprivation Index for patients diagnosed between 1986 and 1995, and the IMD income domain for patients diagnosed after 1996.

3.2.2.2 Health geographies

Two sets of health geographies were used in this research: Primary Care Trusts (PCTs) and Clinical Commissioning Groups (CCGs). Every cancer record was allocated to a PCT and a CCG of residence at diagnosis based on that patients' recorded postcode. The 2011 PCT configuration was used to map each cancer patient to one of the 152 PCTs in England. Using the same procedure, patients were assigned to one of 211 CCGs in England. A map of the CCGs configuration is shown in Figure 3.1. Several boundary changes and mergers have occurred to health geographies in England over the years. CCGs officially replaced PCTs from the 1st April 2013, so that in order to achieve consistency in the geographical units over time, both the PCTs and CCGs boundaries were applied to all the records, based on a combined historic postcode directory covering the entire period of available data [132].

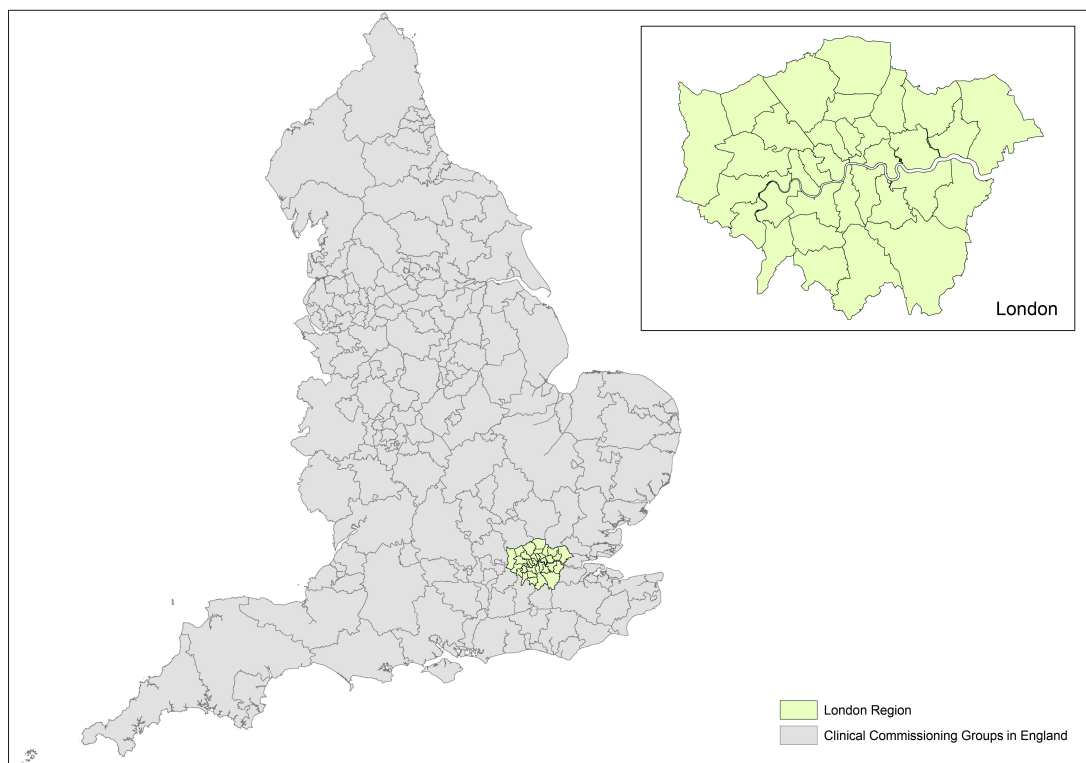


Figure 3.1: Map of the configuration of CCGs in England

3.3 Additional data sources for cancer research

In the last decade many efforts have been made to augment cancer registration data with information contained in several other electronic health databases. The two main sources of data relevant for this research are Hospital Episodes Statistics (HES) [133] and National Bowel Cancer Audit data (NBOCA) [134]. Individual cancer registration records have been linked to HES and NBOCA data, and were available for patients diagnosed between 2006–2013. HES is an administrative database that records information on all admissions, A&E attendance and outpatient appointments at NHS hospitals in England. Data collected includes clinical information about diagnoses and surgical procedures, lengths of hospital stay and information regarding where patients are treated. NBOCA is a collaborative project of clinical audit for bowel cancer in England and Wales, jointly run by the Clinical Effectiveness Unit at the Royal College of Surgeons, the Association of Coloproctology of Great Britain and Ireland (ACPGBI) and NHS Digital. The audit aims to measure the quality of care and outcomes of all patients diagnosed with bowel cancer in England and Wales, and collects data on a range of items which have been identified as good measures of clinical care [135, 136].

3.4 General population life-tables for cancer survival

Mortality hazards from causes other than the cancer of interest are estimated from life tables of the general English population. They are available by single year of age, sex, Government Office Region and deprivation category for each calendar year of death since 1971. These were constructed by the Cancer Survival Group [105, 137] and can be downloaded from the website <https://csg.lshtm.ac.uk/life-tables/> [137].

Chapter 4

An Index of Cancer Survival: a tool for national and local monitoring

"... I believe it is also our job to constantly assess the impact of our activities. One thing I learned from my previous life is this: what gets measured gets done." Dr. Margaret Chan, former WHO Director-General

In this chapter we aimed to summarise and monitor survival for all cancers combined in England at both national and local level (Research Aim 1). We propose a summary survival indicator for all cancers combined based on a three-way standardisation technique. Two sets of weights used in the estimation of the survival index are also presented, depending on the geographical level of the analysis. The implementation of a modelling strategy using excess hazard regression models is described to overcome the additional challenge of estimating the individual cancer survival components needed for the index. Two applications of the cancer survival index are presented: 1) the estimation of 40-year trends in the cancer survival index at one-, five- and ten-years after diagnosis using patients diagnosed with cancer between 1971-2011 in England, i.e. the estimation of the survival index at national level; and 2) the estimation of 16-year trends in the cancer survival index at one-year after diagnosis using patients diagnosed with cancer between 1996-2011 in each of the 211 CCGs in England, i.e. the estimation of the survival index at local level.

4.1 Introduction

Statistics on population-based cancer incidence, mortality and survival are three of the optimal indicators to monitor progress in cancer control efforts [8, 138]. In countries and regions covered by good functioning cancer registries, such statistics are published on a regular basis by government agencies and other official public bodies [7]. In England, several sets of cancer statistics have been published, on a yearly basis, by the Office for National Statistics (ONS) covering the 50-year period since the early 70s [139]. Statistical reports of cancer survival, in particular, have been published in a variety of formats, as for instance, survival for different cancer types and periods of diagnosis, stratified by age groups, regions of residence or socio-economic status [140–142]. The large amount of cancer survival indicators generated over the years have provided crucial baseline evidence to monitor the effectiveness of cancer services in England. Despite the many indicators, national cancer policy-makers, through informal communication with the Cancer Survival Group [143], requested the creation of a new indicator that could summarise the patterns of survival for all cancers combined in ‘one single number’. The indicator was envisioned to become an instrumental monitoring tool at both national and local level. At national level, to act as a surveillance tool of strategic value for the government’s policy. At local level, to serve as a monitoring tool for health service managers, that reflects the outcome of cancer care for patients resident in their areas of catchment.

4.2 Specifications for the development of the index

The new summary indicator of cancer survival was intended to provide a convenient single number that could summarise the overall patterns of cancer survival in each country or region, in each calendar year, for men and women, young and old, and for a wide range of cancers with very different survival. It should reflect that survival for most cancers is either stable or rising steadily from year to year [144]. Patterns of cancer occurrence by age, sex and type of cancer can shift quite quickly over time, especially in small areas. The survival indicator should also reflect real progress (or otherwise) by providing a summary measure of cancer survival that adjusts for any such shifts. It was intended to change only if cancer

survival itself actually changes. The term ‘index of cancer survival’ was chosen for the new all-cancers survival indicator to distinguish it from survival estimates for individual cancers and minimise the risk of misinterpretation.

4.3 Formulation of the index

The index of cancer survival is proposed as a weighted average of cancer survival for every pre-specified combination of sex, age group at diagnosis and type of cancer, as

$$ICS(t) = \sum_{i,j,k} w_{i,j,k} \times S_{i,j,k}(t) \quad (4.1)$$

where $ICS(t)$ is the index of cancer survival at time t after diagnosis, $S_{i,j,k}(t)$ is the cancer survival at time t for every combination of sex i ($i=1,2$), age group j ($j=1,2,\dots,J$) and cancer type k ($k=1,2,\dots,K$), and $w_{i,j,k}$ are the ‘sex-age-cancer’ specific weights, which will be defined in the next sections.

The standard error for the index is given by the weighted average of the standard errors of the sex-age-cancer specific survival, applying the same set of weights

$$se(ICS(t)) = \sqrt{\sum_{i,j,k} w_{i,j,k}^2 \times se(S_{i,j,k}(t))^2} \quad (4.2)$$

Figure 4.1 shows the number of generic combinations of sex, age group and type of cancer needed to estimate the survival index.

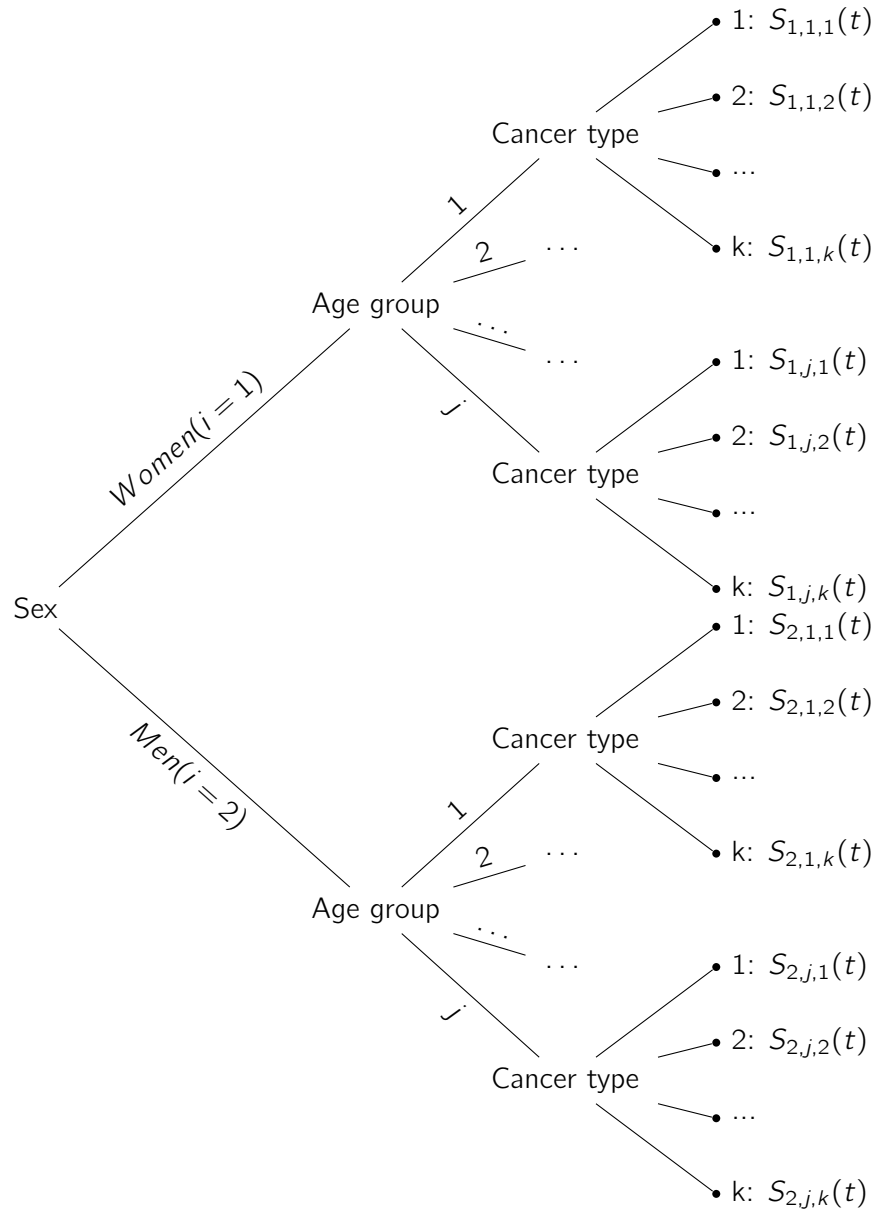


Figure 4.1: Generic combinations needed for the estimation of the index of cancer survival using sex i ($i=1, 2$), age group j ($j=1, 2, \dots, J$) and cancer type k ($k=1, 2, \dots, K$).

4.4 Three-way standardisation technique

The index of cancer survival defined in equation 4.1 is based on a three-way standardisation technique. This approach is an extension to three factors of the classical age-standardisation technique, commonly used when comparing survival between populations [42, 145]. Such survival comparisons can be made within the same population at different time points, or between different populations. Standardisation is needed because for most cancers, the cancer-specific hazard is age-dependent so that when comparing survival between two or more populations, with differential age distributions, the comparisons are misled by the confounding effect of age. Unlike survival estimates that are based on data pooled for all ages, age-standardisation will ensure that any differences in the age distributions between populations, or shifts over time within the same population, will not distort the magnitude of the true differences in survival. In practice, this is achieved by applying a common age distribution, through the use of standard age-weights to all the populations being compared. The most commonly used age-standardisation technique for cancer survival outcomes is the direct method, defined as a weighted average of the age-specific survival, as

$$AS(t) = \sum_{i=1}^n w_i \times S_i(t) \quad (4.3)$$

where $AS(t)$ is the age-standardised cancer survival at time t after diagnosis, $S_i(t)$ is the cancer survival at time t for patients diagnosed in the i^{th} age group, $i=1, \dots, n$, and w_i is the set of age-specific weights from a chosen standard cancer patient population.

The sets of age-specific weights used to standardise cancer survival, correspond to the proportion of patients in each of the defined age groups from a selected standard cancer patient population. This differs from age-standardising incidence rates, which instead use the proportion of individuals from general (not cancer-specific) populations [146, 147]. Several sets of weights have been proposed in the literature for cancer survival age-standardisation, ranging from weights derived from English cancer patient populations [42] to sets derived from cancer cohorts used in international studies of survival [145, 148, 149]. The choice of weights can be seen as arbitrary since the main purpose of standardisation is to access the

differences in survival between populations and not to quantify the actual level of survival in each population being compared. However, it is crucial that the same weights are applied to all the populations being compared to ensure valid comparisons. For all the sets of weights, the sum of weights across all the age groups must add to one (unity) to maintain the numerical consistency of the age-standardised survival estimates. This implies that an estimate of survival is required for every age group for which the set of weights are defined.

The same standardisation principle was applied to implement the index of cancer survival. To make figures from the past comparable with those for today and in the future, and between populations, it is necessary to adjust the survival index for changes over time in the distribution of cancer patients by age, sex, and type of cancer within each population. This is because survival varies widely with all three factors. Overall, cancer survival in a given population can change simply because the distribution of its cancer patients changes, even if survival at each cancer, age and sex combination has not changed. Standardisation for all three factors minimises bias and improves the interpretability of the index.

An alternative approach is briefly discussed at the end of this Chapter.

4.5 Standard weights for the estimation of the index

Similarly to age-standardisation, all the values of the index, past and future, and between populations, need to be adjusted using the same set of weights. In doing so, the cancer survival index will not be affected by changes over time in the proportion of cancers of different lethality in either sex – for example, a reduction in lung cancer or an increase in breast cancer. The index will also be unaffected by a change in the age distribution of newly diagnosed cancer patients, or a shift in the proportion of a given type of cancer between men and women.

Since no adequate sets of weights were available in the literature, we have created two sets of ‘sex-age-cancer’-specific weights to construct the survival index at different levels of geographical aggregation. The first set was created to estimate the survival index for England (i.e. at national level) and the second set to estimate the index for CCGs (i.e. at local level). Both sets of weights were calculated using the same cohort of all adults (aged

15-99 years) diagnosed with cancer in England and Wales during 1996-99 (using data for all years combined). For this cohort of cancer patients, multiple cancers occurring in different anatomical sites or in the same site were excluded to avoid that two or more cancer records for the same patient were counted more than once.

The 'sex-age-cancer'-specific weights were calculated as the proportion of patients in each combination of sex, age group and cancer type. The first set contains 185 combinations of sex (two groups), age (five groups) and cancer type (22 groups) defined as:

Sex

- Male
- Female

Age group

- 15-44 years
- 45-54 years
- 55-64 years
- 65-74 years
- 75-99 years

Cancer type

- | | |
|------------------------|--------------------------------------|
| • Bladder | • Myeloma |
| • Brain | • Non-hodgkin lymphoma (NHL) |
| • Breast (female only) | • Oesophagus |
| • Cervix (female only) | • Ovary (female only) |
| • Colon | • Pancreas |
| • Hodgkin lymphoma | • Prostate (male only) |
| • Kidney | • Rectum |
| • Larynx (male only) | • Stomach |
| • Leukaemia | • Testis (male only) |
| • Lung | • Uterus (female only) |
| • Malignant melanoma | • Other (all other cancers combined) |

The five age groups chosen correspond to the age groups recommended by the International Cancer Survival Standard to age-standardise population-based cancer survival [145]. Although different groupings or more detailed age stratification could have been chosen,

such as 5-year age groups, this would have inevitably increased the number of combinations of sex, age and cancer type due with small number of cases and events, with a potential impact on the estimation of cancer survival in each stratum.

The 22 cancer groups chosen, match the cancer types for which Official Statistics on cancer survival are published in England [150]. These correspond to the 21 most common cancer types (approximately 90% of all the incident cancer cases), plus an additional group combining all the remaining cancer types (mostly rare cancers) into one single group denoted by "Other".

The 'sex-age-cancer'-specific weights for the first set can be found in table 4.1.

Table 4.1: First set of 'sex-age-cancer'-specific weights

Male	Bladder	15-44	0.0006956
Male	Bladder	45-54	0.0023960
Male	Bladder	55-64	0.0066648
Male	Bladder	65-74	0.0135838
Male	Bladder	75-99	0.0157871
Female	Bladder	15-44	0.0002836
Female	Bladder	45-54	0.0007625
Female	Bladder	55-64	0.0019973
Female	Bladder	65-74	0.0044026
Female	Bladder	75-99	0.0077310
Male	Brain	15-44	0.0017485
Male	Brain	45-54	0.0016976
Male	Brain	55-64	0.0021752
Male	Brain	65-74	0.0023839
Male	Brain	75-99	0.0012910
Female	Brain	15-44	0.0012816
Female	Brain	45-54	0.0009552
Female	Brain	55-64	0.0014087
Female	Brain	65-74	0.0018033
Female	Brain	75-99	0.0012548

Sex	Cancer type	Age group	Weight
Female	Breast	15-44	0.0203422
Female	Breast	45-54	0.0394991
Female	Breast	55-64	0.0371995
Female	Breast	65-74	0.0322765
Female	Breast	75-99	0.0379968
Female	Cervix	15-44	0.0059946
Female	Cervix	45-54	0.0023144
Female	Cervix	55-64	0.0017418
Female	Cervix	65-74	0.0017016
Female	Cervix	75-99	0.0020829
Male	Colon	15-44	0.0011023
Male	Colon	45-54	0.0030582
Male	Colon	55-64	0.0079584
Male	Colon	65-74	0.0148707
Male	Colon	75-99	0.0157817
Female	Colon	15-44	0.0010475
Female	Colon	45-54	0.0027906
Female	Colon	55-64	0.0060026
Female	Colon	65-74	0.0121523
Female	Colon	75-99	0.0211395
Male	Hodgkin lymphoma	15-44	0.0019679
Male	Hodgkin lymphoma	45-54	0.0004227
Male	Hodgkin lymphoma	55-64	0.0003866
Male	Hodgkin lymphoma	65-74	0.0003304
Male	Hodgkin lymphoma	75-99	0.0001793
Female	Hodgkin lymphoma	15-44	0.0015639
Female	Hodgkin lymphoma	45-54	0.0002515
Female	Hodgkin lymphoma	55-64	0.0002354
Female	Hodgkin lymphoma	65-74	0.0002408
Female	Hodgkin lymphoma	75-99	0.0002167

Sex	Cancer type	Age group	Weight
Male	Kidney	15-44	0.0007264
Male	Kidney	45-54	0.0019063
Male	Kidney	55-64	0.0032883
Male	Kidney	65-74	0.0045458
Male	Kidney	75-99	0.0034327
Female	Kidney	15-44	0.0004669
Female	Kidney	45-54	0.0009405
Female	Kidney	55-64	0.0016267
Female	Kidney	65-74	0.0025043
Female	Kidney	75-99	0.0027411
Male	Larynx	15-44	0.0002114
Male	Larynx	45-54	0.0010836
Male	Larynx	55-64	0.0020374
Male	Larynx	65-74	0.0025284
Male	Larynx	75-99	0.0017190
Male	Leukaemia	15-44	0.0015304
Male	Leukaemia	45-54	0.0012843
Male	Leukaemia	55-64	0.0024093
Male	Leukaemia	65-74	0.0038448
Male	Leukaemia	75-99	0.0042622
Female	Leukaemia	15-44	0.0010970
Female	Leukaemia	45-54	0.0009766
Female	Leukaemia	55-64	0.0014194
Female	Leukaemia	65-74	0.0024040
Female	Leukaemia	75-99	0.0044227
Male	Lung	15-44	0.0010448
Male	Lung	45-54	0.0060160
Male	Lung	55-64	0.0173830
Male	Lung	65-74	0.0349266
Male	Lung	75-99	0.0314002

Sex	Cancer type	Age group	Weight
Female	Lung	15-44	0.0009351
Female	Lung	45-54	0.0041043
Female	Lung	55-64	0.0089912
Female	Lung	65-74	0.0197014
Female	Lung	75-99	0.0190432
Male	Malignant melanoma	15-44	0.0023505
Male	Malignant melanoma	45-54	0.0019893
Male	Malignant melanoma	55-64	0.0022127
Male	Malignant melanoma	65-74	0.0022715
Male	Malignant melanoma	75-99	0.0018528
Female	Malignant melanoma	15-44	0.0038836
Female	Malignant melanoma	45-54	0.0027304
Female	Malignant melanoma	55-64	0.0024013
Female	Malignant melanoma	65-74	0.0025364
Female	Malignant melanoma	75-99	0.0029337
Male	Myeloma	15-44	0.0002314
Male	Myeloma	45-54	0.0006689
Male	Myeloma	55-64	0.0013993
Male	Myeloma	65-74	0.0023786
Male	Myeloma	75-99	0.0024107
Female	Myeloma	15-44	0.0001405
Female	Myeloma	45-54	0.0004441
Female	Myeloma	55-64	0.0010475
Female	Myeloma	65-74	0.0018501
Female	Myeloma	75-99	0.0028776
Male	NHL	15-44	0.0024802
Male	NHL	45-54	0.0027224
Male	NHL	55-64	0.0039170
Male	NHL	65-74	0.0050621
Male	NHL	75-99	0.0044815

Sex	Cancer type	Age group	Weight
Female	NHL	15-44	0.0015277
Female	NHL	45-54	0.0020508
Female	NHL	55-64	0.0030327
Female	NHL	65-74	0.0042247
Female	NHL	75-99	0.0055906
Male	Oesophagus	15-44	0.0004013
Male	Oesophagus	45-54	0.0017659
Male	Oesophagus	55-64	0.0035438
Male	Oesophagus	65-74	0.0056775
Male	Oesophagus	75-99	0.0057444
Female	Oesophagus	15-44	0.0001204
Female	Oesophagus	45-54	0.0005271
Female	Oesophagus	55-64	0.0012856
Female	Oesophagus	65-74	0.0030501
Female	Oesophagus	75-99	0.0060093
Female	Ovary	15-44	0.0029873
Female	Ovary	45-54	0.0046528
Female	Ovary	55-64	0.0061618
Female	Ovary	65-74	0.0068173
Female	Ovary	75-99	0.0063705
Male	Pancreas	15-44	0.0003064
Male	Pancreas	45-54	0.0010930
Male	Pancreas	55-64	0.0024816
Male	Pancreas	65-74	0.0038889
Male	Pancreas	75-99	0.0040508
Female	Pancreas	15-44	0.000210
Female	Pancreas	45-54	0.0007104
Female	Pancreas	55-64	0.0016963
Female	Pancreas	65-74	0.0034541
Female	Pancreas	75-99	0.0060788

Sex	Cancer type	Age group	Weight
Male	Prostate	15-44	0.0000896
Male	Prostate	45-54	0.0019598
Male	Prostate	55-64	0.013870
Male	Prostate	65-74	0.0372035
Male	Prostate	75-99	0.0457439
Male	Rectum	15-44	0.0007826
Male	Rectum	45-54	0.0030434
Male	Rectum	55-64	0.0073698
Male	Rectum	65-74	0.0116212
Male	Rectum	75-99	0.0102246
Female	Rectum	15-44	0.0007746
Female	Rectum	45-54	0.0020655
Female	Rectum	55-64	0.0036615
Female	Rectum	65-74	0.0064534
Female	Rectum	75-99	0.0101390
Male	Stomach	15-44	0.0005579
Male	Stomach	45-54	0.0017244
Male	Stomach	55-64	0.0044347
Male	Stomach	65-74	0.0092949
Male	Stomach	75-99	0.0101457
Female	Stomach	15-44	0.0003585
Female	Stomach	45-54	0.0006020
Female	Stomach	55-64	0.0015518
Female	Stomach	65-74	0.0037993
Female	Stomach	75-99	0.0078929
Male	Testis	15-44	0.0066206
Male	Testis	45-54	0.0010448
Male	Testis	55-64	0.0003371
Male	Testis	65-74	0.0001391
Male	Testis	75-99	0.0000883

Sex	Cancer type	Age group	Weight
Female	Uterus	15-44	0.0006756
Female	Uterus	45-54	0.0030809
Female	Uterus	55-64	0.0063250
Female	Uterus	65-74	0.0064253
Female	Uterus	75-99	0.0056307
Male	Other	15-44	0.0042260
Male	Other	45-54	0.0061765
Male	Other	55-64	0.0094527
Male	Other	65-74	0.0123022
Male	Other	75-99	0.0112467
Female	Other	15-44	0.0046073
Female	Other	45-54	0.0042314
Female	Other	55-64	0.0057712
Female	Other	65-74	0.0089484
Female	Other	75-99	0.0144118
all	all	all	Sum=1

Similarly to age-standardisation, the sum of weights across all the specified combinations must add to one (unity).

The second set of 'sex-age-cancer'-specific weights contains a total of 35 combinations. The groupings for sex and age at diagnosis are the same as defined in the first set, but the groupings for cancer type are now only defined for:

Cancer-type

- Breast (female only)
- Colorectum (colon and rectum combined)
- Lung
- Other (all other cancers combined)

The 'sex-age-cancer'-specific weights for the second set can be found in table [4.2](#).

Table 4.2: Second set of 'sex-age-cancer'-specific weights

Female	Breast	15-44	0.0203422
Female	Breast	45-54	0.0394991
Female	Breast	55-64	0.0371995
Female	Breast	65-74	0.0322765
Female	Breast	75-99	0.0379968
Female	Colorectum	15-44	0.0018220
Female	Colorectum	45-54	0.0048561
Female	Colorectum	55-64	0.0096641
Female	Colorectum	65-74	0.0186058
Female	Colorectum	75-99	0.0312785
Male	Colorectum	15-44	0.0018849
Male	Colorectum	45-54	0.0061016
Male	Colorectum	55-64	0.0153282
Male	Colorectum	65-74	0.0264920
Male	Colorectum	75-99	0.0260063
Female	Lung	15-44	0.0009351
Female	Lung	45-54	0.0041043
Female	Lung	55-64	0.0089912
Female	Lung	65-74	0.0197014
Female	Lung	75-99	0.0190432
Male	Lung	15-44	0.0010448
Male	Lung	45-54	0.0060160
Male	Lung	55-64	0.0173830
Male	Lung	65-74	0.0349266

Sex	Cancer type	Age group	Weight
Male	Lung	75-99	0.0314002
Female	Other	15-44	0.0251983
Female	Other	45-54	0.0252304
Female	Other	55-64	0.0377025
Female	Other	65-74	0.0541625
Female	Other	75-99	0.0762451
Male	Other	15-44	0.0241442
Male	Other	45-54	0.0279354
Male	Other	55-64	0.0586106
Male	Other	65-74	0.1054354
Male	Other	75-99	0.1124360
all	all	all	Sum=1

4.6 Application: Index of cancer survival for England and CCGs

For England, an index of cancer survival was estimated for all adult patients (aged 15-99 years) diagnosed between 1971 and 2011 and followed up to the end of 2012. The index was estimated at one-, five-, and ten-years after diagnosis for six selected periods: 1971-72, 1980-81, 1990-91, 2000-01, 2005-2006 and 2010-11.

For CCGs, an index of cancer survival was estimated for each of the 211 CCGs, including all adult patients (aged 15-99 years) diagnosed between 1996-2011. The index was estimated at one-year after diagnosis for every calendar year between 1996 and 2011.

For both indexes, multiple cancers occurring in different anatomical sites or in the same site were excluded to avoid that two or more cancer records for the same patient contributed to the survival analysis.

4.6.1 Estimation of the individual index components

Constructing the indexes of cancer survival requires the estimation of several individual components. These are the estimates of cancer survival for each combination of sex, age group at diagnosis and cancer type. The total number of combinations and thus survival estimates needed for each index, depends on the groupings chosen for the three factors. To estimate the index for England the groupings defined for the first set of weights were used (Table 4.1). These correspond to a total of 185 'sex-age-cancer' combinations and use as cancer-type groupings the 21 most common cancer types as published by official statistics on cancer outcomes indicators [141]. For this set of weights all the other remaining cancers types are combined into one group called 'other cancers'. To estimate the index for each CCG, the groupings that define the second set of weights were chosen (Table 4.2). These groupings add up to a total of 35 'sex-age-cancer' combinations, less than the 185 combinations used in the first set of weights. This reduction in the number of combinations, tries to minimise the potential number of missing 'sex-age-cancer' survival estimates, that can arise due to the smaller number of cases and events occurring in each of the 211 CCGs compared to the whole of England.

4.6.2 Developing a modelling strategy for the estimation of net survival

Flexible parametric excess hazard regression models (as defined in equation 2.19) were used to estimate the required net survival components to construct the two indexes using the user-written command *stpm2* [100] in Stata software [103].

For both indexes, models were fitted separately for men and women, and for each cancer type (some of which are gender-specific as shown in tables 4.1 and 4.2). This added to a total of 37 regression models that needed to be fitted for the England index and to 1,477 models for the CCG index (modelling 7 gender-specific cancer groups for 211 CCGs).

This large number of models (1,514 in total for the two indexes) made model fitting a cumbersome task to be undertaken. Making it in particular challenging to use classical model selection approaches to fit each model ‘manually’ with stepwise selection procedures [151]. After some consideration, we have decided that the best approach to overcome this challenge was to develop a ‘semi-automated’ modelling strategy. For this purpose, we wrote a Stata algorithm that fits up to eight candidate flexible parametric survival models using *stpm2* [100]. All the models include age and year of diagnosis as main effects to enable the estimation of net survival by age group and year (or period) of diagnosis. For each fitted model, the algorithm tests non-linear and non-proportional effects, and interaction terms and chooses the best fitting model based on the Akaike Information Criterion (AIC) [152]. After the best fitting model is chosen, the algorithm proceeds to the post-estimation prediction of net survival for all the components of the indexes.

Complete details regarding the development and implementation of this algorithm, including model set-up, choice of splines, model selection and post-estimation of net survival are described in the next few sections.

4.6.3 Model set-up

Age and year of diagnosis were included in all the models as main effects. The inclusion of age at diagnosis in the models allowed for the non-informative censoring process to be taken into account in the estimation of net survival as mentioned in Chapter 2. Restricted cubic

splines were used to model any potential non-linear effects of age and year of diagnosis, and of the interaction between age and year. The baseline log-cumulative excess hazard was also modelled using restricted cubic splines. To account for potential non-proportionality of the excess hazards over time, interactions were considered between follow-up time and age, year, and the interaction between age and year.

Background mortality rates were obtained from English life tables stratified by single year of age, sex, region of residence and deprivation category, for every calendar year. For the England index, national or regional life-tables were used for the 2.8% of patients who could not be assigned to a specific deprivation category or region; almost all of these patients were diagnosed in the 1970s (85%) or 1980s (14%).

Seven candidate models were specified *a priori* on the log-cumulative excess hazard scale for both indexes. An additional and simpler model was defined for the CCG index. These candidate models were defined based on our previous experience in modelling cancer survival. The models can be formally written as:

Model 1 (M1) Non-linear and non-proportional effects of age and year of diagnosis, and a non-linear and non-proportional interaction between age and year of diagnosis

$$\begin{aligned} \ln(H_{net}(t|age, year)) = & s_0(\ln(t)|kn_0) + f(age) + s(\ln(t)|kn_{age}).f(age) + g(year) \\ & + s(\ln(t)|kn_{year}).g(year) + l(age.year) \\ & + s(\ln(t)|kn_{age.year}).l(age.year) \end{aligned} \quad (4.4)$$

Model 2 (M2) Non-linear and non-proportional effects of age and year of diagnosis, and a non-linear interaction between age and year of diagnosis

$$\begin{aligned} \ln(H_{net}(t|age, year)) = & s_0(\ln(t)|kn_0) + f(age) + s(\ln(t)|kn_{age}).f(age) + g(year) \\ & + s(\ln(t)|kn_{year}).g(year) + l(age.year) \end{aligned} \quad (4.5)$$

Model 3 (M3) Non-linear and non-proportional effects of age and year of diagnosis

$$\begin{aligned} \ln(H_{net}(t|age, year)) = & s_0(\ln(t)|kn_0) + f(age) + s(\ln(t)|kn_{age}).f(age) + g(year) \\ & + s(\ln(t)|kn_{year}).g(year) \end{aligned} \quad (4.6)$$

Model 4 (M4) Non-linear and non-proportional effects of age and non-linear year of diagnosis

$$\ln(H_{net}(t|age, year)) = s_0(\ln(t)|kn_0) + f(age) + s(\ln(t)|kn_{age}).f(age) + g(year) \quad (4.7)$$

Model 5 (M5) Non-linear and non-proportional effect of age and a linear and proportional year of diagnosis

$$\ln(H_{net}(t|age, year)) = s_0(\ln(t)|kn_0) + f(age) + s(\ln(t)|kn_{age}).f(age) + year \quad (4.8)$$

Model 6 (M6) Linear effect of age and year of diagnosis and non-proportional effect of age

$$\ln(H_{net}(t|age, year)) = s_0(\ln(t)|kn_0) + age + s(\ln(t)|kn_{age}).age + year \quad (4.9)$$

Model 7 (M7) Non-linear effects of age and year of diagnosis, and a non-proportional effect of year of diagnosis and a non-proportional interaction between age and year of diagnosis

$$\begin{aligned} \ln(H_{net}(t|age, year)) = & s_0(\ln(t)|kn_0) + f(age) + g(year) + s(\ln(t)|kn_{year}).g(year) \\ & + s(\ln(t)|kn_{age.year}).(age.year) \end{aligned} \quad (4.10)$$

And the additional model for the CCG index,

Model 8 (M8) Non-linear effects of age and year of diagnosis, and a linear and non-proportional effect of the interaction between age and year of diagnosis

$$\begin{aligned} \ln(H_{net}(t|age, year)) = & s_0(\ln(t)|kn_0) + f(age) + g(year) + age \cdot year + \\ & + s(\ln(t)|kn_{age \cdot year}) \cdot (age \cdot year) \end{aligned} \quad (4.11)$$

where,

- $\ln(H_{net}(t|age, year))$ is the log-cumulative excess hazard to be modelled.
- $s_0(\ln(t)|kn_0)$ is the non-linear effect of the baseline log-cumulative excess hazard, $\ln(t)$ is the logarithm of time after diagnosis and kn_0 is the number of knots for the spline.
- $f(age)$ is the non-linear effect of age at diagnosis.
- $s(\ln(t)|kn_{age}) \cdot f(age)$ is the non-proportional (and non-linear) effect of age at diagnosis.
- $g(year)$ is the non-linear effect of year of diagnosis.
- $s(\ln(t)|kn_{year}) \cdot g(year)$ is the non-proportional (and non-linear) effect of year of diagnosis.
- $l(age \cdot year)$ is the non-linear effect of the interaction between age and year of diagnosis.
- $s(\ln(t)|kn_{age \cdot year}) \cdot l(age \cdot year)$ is the non-proportional (and non-linear) effect of the interaction between age and year of diagnosis.
- age is the linear effect of age at diagnosis.
- $year$ is the linear effect of year of diagnosis.
- $age \cdot year$ is the linear effect of the interaction between age and year at diagnosis.
- $s(\ln(t)|kn_{age}) \cdot age$ is the non-proportional (and linear) effect of age at diagnosis, kn_{age} is the number of knots for the splines on age at diagnosis.
- $s(\ln(t)|kn_{age \cdot year}) \cdot (age \cdot year)$ is the non-proportional (and linear) effect of the interaction between age and year of diagnosis, $kn_{age \cdot year}$ is the number of knots for the splines on the interaction.

4.6.4 Selecting the number and location of spline knots

The Stata command *rcs*gen [153] was used to create restricted cubic splines to model the non-linear effects of age and year of diagnosis, of the interaction between age and year, and of the baseline log-cumulative excess hazard function. These types of splines are restricted by construct to be linear before the first knot and after the last knot. The knots that can be specified are called ‘internal knots’ and are equal to the number of degrees of freedom minus 1. They can be defined in two ways: either by setting the number of knots so that the exact knot location will be at the corresponding percentiles of the distribution, or by choosing the exact knot location at given values of the distribution [97].

For both the national and local indexes, three knots (with four degrees of freedom) were chosen to model the baseline log-cumulative excess hazard. The knots were defined at the 25th, 50th and 75th percentiles of the distribution. For both the non-linear effect of year of diagnosis and the interaction term between age and year, two knots (with three degrees of freedom) were defined at the 33rd and 66th percentiles of the distribution.

The knot locations for the splines modelling the non-linear effect of age at diagnosis were chosen separately for each cancer type. This is because the effect of age on the excess hazard is different according to the type of cancer and therefore different shapes for the splines modelling those effects are needed. Exact knot locations were chosen based on the histograms of the age distribution of each cancer and on previous knowledge about the general effect of age on the cancer-specific excess hazards.

The knot locations chosen for the England index are,

Cancer type	Knot location (age at diagnosis)
testis	15 35 99
leukeamia	15 45 75 99
hodgkin lymphoma	15 25 65 99
cervix, melanoma	15 35 65 99
brain, ovary	15 40 65 99
breast, colon, uterus, NHL	15 50 70 99
(bladder, kidney, larynx, myeloma, oesophagus prostate, rectum, stomach, other cancers)	15 65 99
lung, pancreas	15 40 65 75 99

Table 4.3: Location of knots for the spline on age at diagnosis for the England index

The knot locations chosen for the CCG index are,

Cancer type	Knot location (age at diagnosis)
breast, colorectum	15 50 70 99
lung, other cancers	15 65 99

Table 4.4: Location of knots for the spline on age at diagnosis for the CCG index

4.6.5 Model selection using the Akaike Information Criterion

The model set-up specified in section 4.6.3 was implemented in Stata for both indexes. For the England index, seven models were run for each of the 37 combinations of sex and cancer grouping, adding to a total of 259 models. For the CCG index, eight models were run for each of the 1,477 combinations of CCG, sex and cancer grouping, adding to a total of 11,816 fitted models. The best-fitting model was chosen as the one with the smallest value of AIC for each of the relevant combinations of 'England-sex-cancer-type' or 'CCG-sex-cancer-type'.

4.6.6 Post-estimation of net survival

After the best fitting models were selected, the Stata *post-estimation* command *predictnl*, with the option *meansurv*, was used to calculate net survival for every component of the index. This command is compatible with the *stpm2* [100] post-estimation in Stata and it implements the following steps:

Step 1 A cumulative survival function is derived for each observation in the dataset using the Maximum Likelihood (ML) parameter estimates and the expression for the survivor function defined in equation (2.2).

Step 2 The individual survival functions are averaged to derive net survival for each combination of sex, age group and cancer-type and for each year (or period) of diagnosis. The averaging includes all the patients that fall within each of the relevant combinations.

Step 3 Standard errors and 95% confidence intervals are obtained for each estimated net survival function using the Delta method [154].

Stata code for the implementation of the national and the local indexes is provided in Appendix A.

4.6.7 Dealing with model non-convergence

For every cancer-sex combination for which none of the candidate models converged, models were refitted individually, adjusting the number and location of spline knots for each of the effects being modelled. If models still did not converge, for a particular cancer-sex combination in a given CCG, the missing estimate was replaced by the equivalent estimate for England for the same cancer-sex combination. For models that did converge, but for which the post-estimation did not provide an estimate of net survival for a specific combination of sex, age group and cancer, the post-estimation was re-run using a merged age group that was then used to replace the missing age group estimate. For example, if an estimate was missing for the 15-44 age group, the post-estimation was re-run using a new

age group '15-54' (merging the '15-44' age group with the adjacent '45-54' age group) and the re-estimated survival was used for the missing age group.

4.6.8 Combining the individual components of the index

The index of cancer survival is calculated by combining the individual net survival components through the weighted average formula defined in equation (4.1). The indexes can be calculated for several years (or periods) of diagnosis and for specific times after diagnosis, such as one-, five- or ten- years after diagnosis. Standard errors can be calculated for each of the estimated indexes using equation (4.2).

4.6.9 Results 1: Index of cancer survival for England [1]

The results of the index for England have been peer-reviewed and published in the *The Lancet* [1]. This is the first research publication of this PhD. The original article is inserted at the end of this section. The article also presents trends in age-sex adjusted survival for each cancer and analysed the absolute change (%) in the age gap in survival since 1971. For comparative purposes, equivalent cancer patients data from the Welsh Cancer Intelligence & Surveillance Unit, was also used to construct an index of cancer survival for Wales. A summary of model convergence and of the main index estimates is presented below.

Table 4.5 presents an overview of the patterns of model convergence for the 7 models that were fitted for each cancer-sex combination for the England index. The models that converged for each combination are indicated by the symbol 'x' and the last column indicates the best fitting model selected based on the smallest value of the AIC.

Table 4.5: Models run for each cancer-sex combination (M1-M7, in the same order as specified in section 4.6.3); converged models marked with 'x'; and best model selected

Cancer type	Sex	M1	M2	M3	M4	M5	M6	M7	Model selected
hodgkin lymphoma	male	.	x	x	x	x	x	x	M2
hodgkin lymphoma	female	x	x	x	x	x	x	x	M1
non-hodgkin lymphoma	male	.	x	.	x	x	x	x	M2
non-hodgkin lymphoma	female	.	x	x	x	x	x	x	M2
bladder	male	.	x	.	x	x	x	x	M2
bladder	female	.	.	.	x	x	x	x	M4
brain	male	x	x	x	x	x	x	x	M1
brain	female	x	x	x	x	x	x	x	M1
breast	female	x	x	x	x	x	x	x	M1
cervix	female	.	x	x	x	x	x	x	M2
colon	male	.	x	x	x	x	x	x	M2
colon	female	x	x	x	x	x	x	x	M1
kidney	male	x	x	x	x	x	x	x	M1
kidney	female	x	x	x	x	x	x	x	M1
larynx	male	x	x	M6
leukaemia	male	.	x	x	.	.	x	x	M2
leukaemia	female	x	x	x	x	x	x	x	M1
lung	male	x	M7
lung	female	.	x	x	x	x	.	x	M2
melanoma	male	x	x	x	x	x	x	x	M1
melanoma	female	.	x	x	x	x	x	x	M2
myeloma	male	x	x	x	x	x	x	x	M1
myeloma	female	x	x	x	M1
oesophagus	male	.	x	x	.	.	x	x	M2
oesophagus	female	x	x	x	x	x	x	x	M1
others	male	x	x	x	x	x	x	x	M1
others	female	x	x	x	x	x	x	x	M1
ovary	female	.	x	x	x	x	x	x	M2
pancreas	male	x	x	M7
pancreas	female	x	x	x	x	x	x	x	M1
prostate	male	.	x	x	.	x	x	x	M2
rectum	male	x	.	M6
rectum	female	x	x	M6
stomach	male	.	x	x	x	x	x	x	M2
stomach	female	.	x	x	x	x	x	x	M2
testis	male	x	.	M6
uterus	female	x	x	x	x	x	x	x	M1

Table 4.6 provides a summary of the number of times each of the 7 candidate models were selected for all the cancer-sex combinations as shown in Table 4.5. Model 1 and Model 2, the two most complex candidate models, were selected as best fitting models for 81% of cancer-sex combinations, whilst Model 6, the simplest model, was the third most selected for 11% of the combinations. This mostly occurred when Model 6 was the only, or one of the few models, to converge for that particular cancer-sex combination.

Candidate model	No. of times selected	%
Model 1	16	43.2
Model 2	14	37.8
Model 3	0	0.0
Model 4	1	2.7
Model 5	0	0.0
Model 6	4	10.8
Model 7	2	5.4

Table 4.6: England index: summary of best fitting models for all the cancer-sex combinations

For every cancer-sex combination, there was at least one model (out of the seven candidate models) that converged and that could be used in the post-estimation procedure. No missing estimates occurred during the estimation of net survival for each of the components.

The index of cancer survival for England increased substantially at one-, five- and ten-years after diagnosis between 1971 and 2011 (Figure 4.2). The index was estimated at 50% at one year after diagnosis for patients diagnosed in 1971-72. For patients diagnosed during 2005-2006, the index was 50% at five years after diagnosis, and for patients diagnosed during 2010-2011, we predicted that the index would reach 50% at ten years after diagnosis. Estimates are shown as percentages (0-100) since this is the most common scale cancer survival estimates are presented but they refer to survival probabilities taking values between 0 and 1.

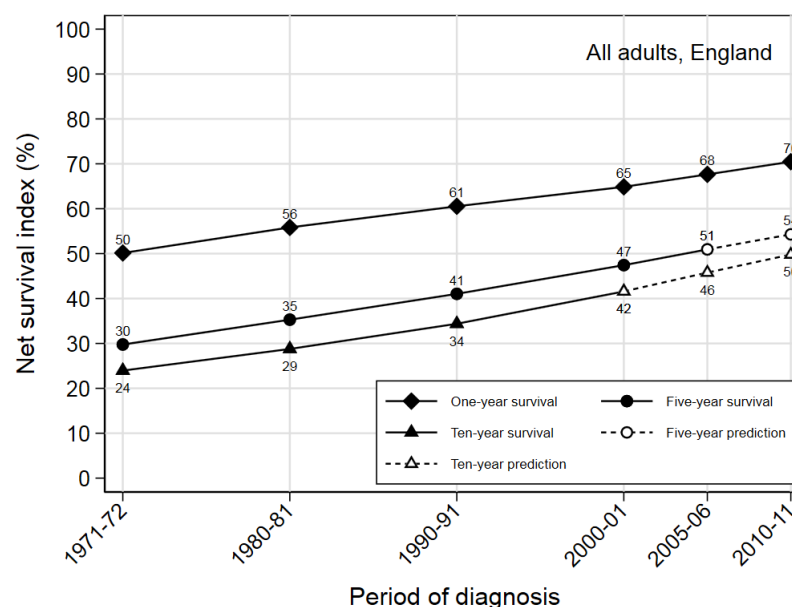


Figure 4.2: Trends in the index of net survival for all cancers combined in England

Figure 4.3 presents scatter plots of 1-year, 5-year and 10-year net survival adjusted for age and sex for each cancer in 2010-11, and the absolute change since 1971 for all adult patients (aged 15-99) diagnosed in England. The absolute change was calculated as the simple arithmetic difference between net survival in 2010-11 and the survival in 1971.

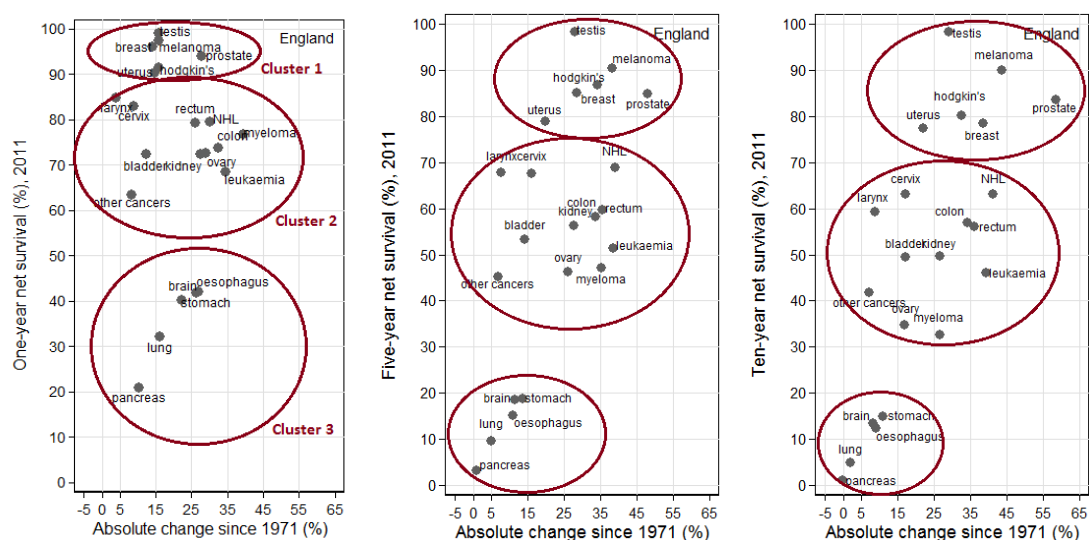


Figure 4.3: Net survival adjusted for age and sex for each cancer in 2010-11, and absolute change since 1971, England: 1, 5, and 10 years after diagnosis

Survival for both sexes combined varied widely for different cancers, with the most recent predicted 10-year net survival adjusted for age and sex ranging from only 1,1% for pancreatic cancer to 98,2% for testicular cancer.

We were able to group visually the 21 most common cancers into three broad clusters on the basis of their survival level. These clusters were identifiable as early as 1 year after diagnosis and they were consistent at 5 and 10 years after diagnosis. The first cluster includes cancers with a good prognosis: survival is very high in 2010-11 after a large increase in survival since 1971-72 at 1, 5, and 10 years. It includes cancers of the breast, prostate, testis, and uterus, and melanoma and Hodgkin's disease. 1-year survival seems to have reached a ceiling for most of these cancers, but 5- and 10-year survival is still much lower than 1 year for cancers of the breast, uterus and Hodgkin's disease. The second cluster includes cancers with a moderate level of survival (64-84%) in 2010-11. This cluster includes a mix of cancers for which survival either has remained moderate since the 1970s (larynx, cervix, bladder and ovary), or moderate levels of survival in 2011 that are the result of large improvements during the past 40 years (rectum, colon, kidney, non-Hodgkin lymphoma, multiple myeloma, and leukaemia). The third cluster includes cancers with very low survival in 2010-11, for which little or no improvement has occurred in the past 40 years: this group consists of cancers of the brain, stomach, lung, oesophagus, and pancreas.

These findings support substantial increases in both short-term and long-term net survival from all cancers combined in England. They also highlight individual cancer types for which prognosis remains very poor, at 5- and 10-years after diagnosis particularly. The index of net survival provides one convenient number that summarises the overall patterns of cancer survival in any one population, in each calendar period, for young and old men and women and for a wide range of cancers with very disparate survival. It was designed as a public health measure to help assess progress in the overall effectiveness of the health system in diagnosis and management of patients with cancer. The index should nevertheless be interpreted in conjunction with other information available in the population for which the index has been prepared. It should be seen as a guide to raise questions about the potential for improvement.

4.6.9.1 Research publication 1

Title: '40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study'

Authors: Manuela Quaresma, Michel P Coleman and Bernard Rachet.

Peer-reviewed and published in *The Lancet*. The final published article inserted from next page.

Copyright © Quaresma et al. Open Access article distributed under the terms of CC BY.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	199304	Title	Ms
First Name(s)	Manuela		
Surname/Family Name	Quaresma		
Thesis Title	Population-based cancer survival at small area level: methodological developments		
Primary Supervisor	Professor Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	The Lancet		
When was the work published?	2015		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	Published within PhD registration period		
Have you retained the copyright for the work?*	No	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	MQ did the analysis. MQ and BR designed the analytic strategies and constructed the indexes. MQ, MPC, and BR wrote the article and interpreted the findings.
--	--

SECTION E

Student Signature	
Date	20/11/2019

Supervisor Signature	
Date	20/11/2019



40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study



Manuela Quaresma, Michel P Coleman, Bernard Rachet

Lancet 2015; 385: 1206–18

Published Online

December 3, 2014

[http://dx.doi.org/10.1016/S0140-6736\(14\)61396-9](http://dx.doi.org/10.1016/S0140-6736(14)61396-9)

50140-6736(14)61396-9

This online publication has been corrected. The corrected version first appeared at thelancet.com on March 27, 2015

See [Comment](#) page 1162

Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

(M Quaresma Licenciatura, Prof M P Coleman FFPH, B Rachet FFPH)

Correspondence to: Ms Manuela Quaresma, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK
manuela.quaresma@lshtm.ac.uk

Summary

Background Assessment of progress in cancer control at the population level is increasingly important. Population-based survival trends provide a key insight into the overall effectiveness of the health system, alongside trends in incidence and mortality. For this purpose, we aimed to provide a unique measure of cancer survival.

Methods In this observational study, we analysed trends in survival with population-based data for 7·2 million adults diagnosed with a first, primary, invasive malignancy in England and Wales during 1971–2011 and followed up to the end of 2012. We constructed a survival index for all cancers combined using data from the National Cancer Registry and the Welsh Cancer Intelligence and Surveillance Unit. The index is designed to be independent of changes in the age distribution of patients with cancer and of changes in the proportion of lethal cancers in each sex. We analysed trends in the cancer survival index at 1, 5, and 10 years after diagnosis for the selected periods 1971–72, 1980–81, 1990–91, 2000–01, 2005–06, and 2010–11. We also estimated trends in age-sex-adjusted survival for each cancer. We define the difference in net survival between the oldest (75–99 years) and youngest (15–44 years) patients as the age gap in survival. We evaluated the absolute change (%) in the age gap since 1971.

Findings The overall index of net survival increased substantially during the 40-year period 1971–2011, both in England and in Wales. For patients diagnosed in 1971–72, the index of net survival was 50% at 1 year after diagnosis. 40 years later, the same value of 50% was predicted at 10 years after diagnosis. The average 10% survival advantage for women persisted throughout this period. Predicted 10-year net survival adjusted for age and sex for patients diagnosed between 2010 and 2011 ranged from 1·1% for pancreatic cancer to 98·2% for testicular cancer. Net survival for the oldest patients (75–99 years) was persistently lower than for the youngest (15–44 years), even after adjustment for the much higher mortality from causes other than cancer in elderly people.

Interpretation These findings support substantial increases in both short-term and long-term net survival from all cancers combined in both England and Wales. The net survival index provides a convenient, single number that summarises the overall patterns of cancer survival in any one population, in each calendar period, for young and old men and women and for a wide range of cancers with very disparate survival. The persistent sex difference is partly due to a more favourable cancer distribution in women than men. The very wide differences in survival for different cancers, and the persistent age gap in survival, suggest the need for renewed efforts to improve cancer outcomes. Future monitoring of the cancer survival index will not be possible unless the current crisis of public concern about sharing of individual data for public health research can be resolved.

Funding Cancer Research UK.

Copyright © Quaresma et al. Open Access article distributed under the terms of CC BY.

Introduction

Cancer is an increasing public health concern, shown by substantial investments in human and financial resources for cancer management since the late 1990s. Health policy measures have focused on improvement of the organisation and delivery of services for prevention, diagnosis, and treatment. Research has provided the evidence base for these policies and is increasingly used to assess their effect.^{1–7} The assessment of progress in cancer control has become crucial. Population-based cancer survival trends provide a key insight into the overall effectiveness of the health system, alongside incidence and mortality.⁸

In this population-based survival study, we analysed cancer survival trends during the past four decades in England and Wales using two metrics: an index of survival for all cancers combined, and survival for each cancer, adjusted for age and sex. The all-cancers survival index was designed to provide one summary measure of cancer survival that can be monitored over time to show the overall progress in the effectiveness of the health-care system. It was also designed to support assessment of the effect of earlier diagnosis, which is a key component of the National Awareness and Early Diagnosis Initiative.^{9–11} Trends in survival for individual cancers will underline those cancer types for which

there has been progress and those for which prognosis has remained poor.

Methods

Study design

Survival varies very widely with the age and sex of a patient with cancer and with the type of cancer. The frequency of different cancers is also changing over time: some cancers with poor prognosis, such as stomach and lung cancer, have become less common, whereas breast cancer in women, for which survival has been improving, has become more common. These trends can differ between the sexes: lung cancer has become much less common in men, but more common in women. The age profile of patients with cancer also changes over time, and these trends can differ between cancers. To enable valid assessment of survival trends for all cancers combined, the survival index must therefore take account of changes over time in the distribution of age, sex, and cancer type in all patients with cancer, especially over periods as long as 40 years. Similarly, trends in survival for each cancer must be adjusted for changes over time in the age (and sex) profile of patients with cancer.

Data sources

We examined survival trends in 7176795 adults (aged 15–99 years) diagnosed with a first, primary, invasive malignancy in England and Wales during 1971–2011, and followed up to Dec 31, 2012 (table 1). Data for England were obtained from the National Cancer Registry at the Office for National Statistics¹² and for Wales from the Welsh Cancer Intelligence and Surveillance Unit. Patients diagnosed with a malignancy of the skin other than melanoma were excluded. Since 1971, the National Health Service Central Register has routinely updated these individual cancer records with information about each patient's vital status (alive, emigrated, dead, or not traced). The vital status at Dec 31, 2012, was known for 98·4% of these patients. During the 41-year period, 4·3% of all cancer registrations were for the patient's second-order or higher-order tumour: in the analyses for all cancers combined, the higher-order cancers were not included.

Statistical analysis

The all-cancers survival index was constructed as a weighted average of the survival estimates for every combination of age group at diagnosis (15–44, 45–54,

	ICD-10 code*	England				Wales			
		Women		Men		Women		Men	
		Number	%	Number	%	Number	%	Number	%
Oesophagus	C15	67 474	2·0%	106 793	3·1%	4953	2·3%	6857	3·1%
Stomach	C16	115 294	3·4%	194 333	5·7%	8627	4·0%	14 299	6·5%
Colon	C18	292 352	8·7%	271 220	8·0%	17 711	8·3%	17 736	8·1%
Rectum	C19–C21	143 610	4·3%	204 363	6·0%	9731	4·5%	14 358	6·6%
Pancreas	C25	92 631	2·8%	93 450	2·7%	5868	2·7%	6014	2·7%
Larynx (men)	C32	52 618	1·5%	3529	1·6%
Lung	C33, C34	349 711	10·5%	751 958	22·1%	21 027	9·8%	45 601	20·8%
Melanoma	C43	97 627	2·9%	72 743	2·1%	5429	2·5%	4372	2·0%
Breast (women)	C50	1 039 609	31·1%	65 370	30·6%
Cervix	C53	117 404	3·5%	8272	3·9%
Uterus	C54, C55	160 539	4·8%	10 836	5·1%
Ovary	C56, C57·0–7	172 400	5·2%	11 051	5·2%
Prostate	C61	638 111	18·8%	41 559	19·0%
Testis	C62	48 031	1·4%	2743	1·3%
Kidney	C64–C66, C68	53 197	1·6%	89 986	2·6%	3431	1·6%	5804	2·6%
Bladder	C67	90 204	2·7%	239 621	7·0%	5897	2·8%	15 962	7·3%
Brain	C71	41 952	1·3%	59 192	1·7%	2832	1·3%	3786	1·7%
Hodgkin's disease	C81	19 114	0·6%	26 714	0·8%	1145	0·5%	1675	0·8%
Non-Hodgkin lymphoma	C82–C85	99 752	3·0%	114 269	3·4%	5630	2·6%	6320	2·9%
Myeloma	C90	43 446	1·3%	48 136	1·4%	2805	1·3%	3041	1·4%
Leukaemia	C91–C95	70 760	2·1%	92 917	2·7%	4686	2·2%	6112	2·8%
Other cancers†	..	275 408	8·2%	296 794	8·7%	18 624	8·7%	19 369	8·8%
Total	..	3 342 484	100·0%	3 401 249	100·0%	213 925	100·0%	219 137	100·0%

*Tenth revision of the International Classification of Diseases (ICD): malignancies were initially coded according to the ICD revision in use during the year of diagnosis—ie, ICD 8 (1971–78), 9 (1979–95), or 10 (1996–). †Other cancers: all other malignant tumours are combined; they also include laryngeal cancer in women and breast cancer in men.

Table 1: Number of patients (aged 15–99 years) included in analyses in England and Wales diagnosed from 1971 to 2011 and followed up to 2012, by sex and type of malignancy

55–64, 65–74, and 75–99 years), sex (male and female), and type of cancer (the 21 most common malignancies are shown in table 1 and all other malignant tumours are combined). The weights used were the proportion of patients with cancer diagnosed in England and Wales during 1996–99 in each of the 185 combinations of age group, sex, and type of cancer. We also constructed the all-cancers survival index separately for males and females and estimated survival adjusted for age and sex by cancer.

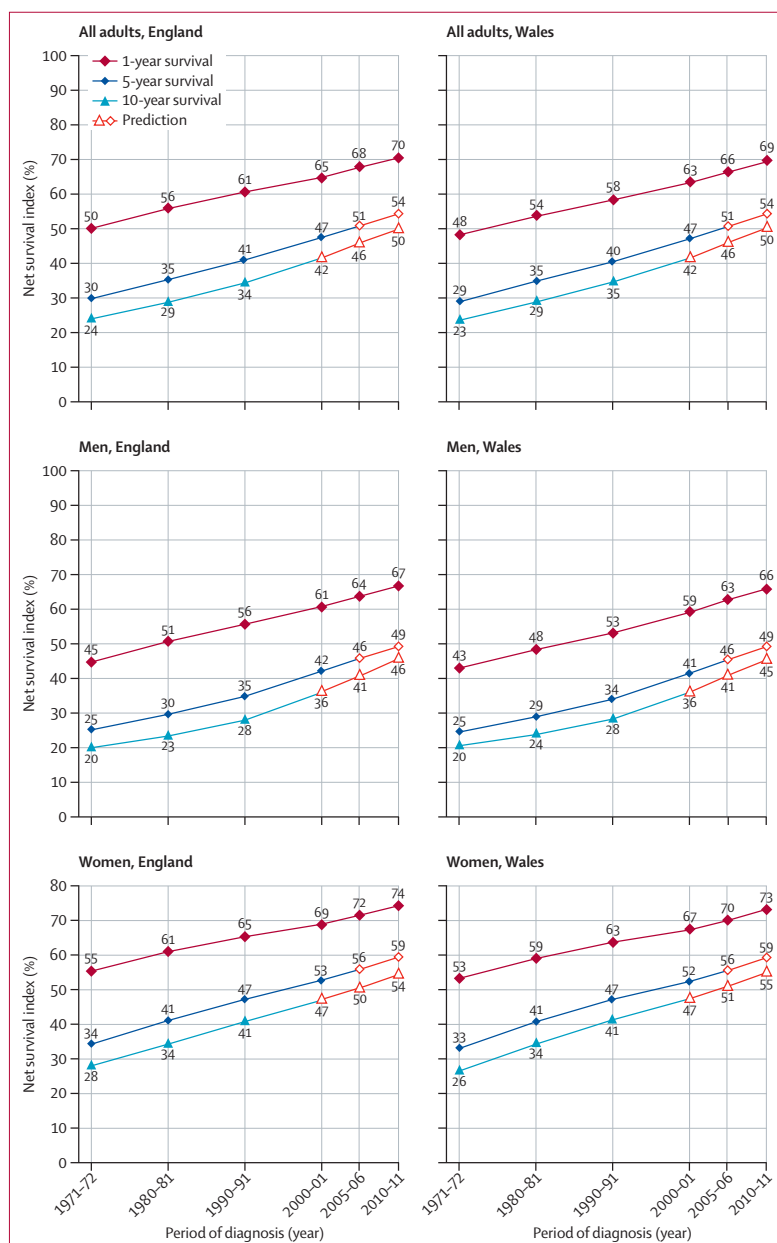


Figure 1: Trends in the index of net survival for all cancers combined, for England and for Wales: all adults (15–99 years), men, and women, selected periods during 1971–2011

Net survival was used as the cancer survival measure for each component of the indexes. Net survival quantifies the survival after taking account of death from other causes (background mortality). All patients were allocated a deprivation category defined according to their Lower Super Output Area (mean population about 1500) of residence at the time of cancer diagnosis. Life-tables were used to take account of the wide variation in background mortality by age, sex, deprivation, region, and over time. For this study, separate life-tables were created for England and Wales by single year of age, sex, deprivation category, and (in England) region of residence, for every calendar year between 1971 and 2012.¹³ National or regional life-tables were used for the 2.8% of patients diagnosed in England (2.6% in Wales) who could not be assigned to a specific deprivation category or (in England) region; almost all of these patients were diagnosed in the 1970s (85% in England, 55% in Wales) or 1980s (14% England, 44% Wales).

We used flexible multivariable parametric excess hazard models^{14,15} to estimate net survival up to 10 years after diagnosis for each nation, and for each stratum defined by cancer, sex, age group, and calendar period. The models included age and year of diagnosis as main effects, modelled on a continuous scale with restricted cubic splines, to account for potential non-linear excess (cancer-related) hazards. Interactions between age and year of diagnosis, year of diagnosis and follow-up time, and age and follow-up time were assessed to deal with potential variation of the excess hazard with time since diagnosis. The best-fitting models were chosen as those with the smallest Akaike Information Criterion.¹⁶ Net survival curves were estimated for each individual from these models according to their age and year of diagnosis. We obtained net survival estimates for each cancer and sex by averaging of individual net survival curves, over all ages and years of diagnosis within each age group and calendar period. In view of the fact that the models included the year of diagnosis as a continuous variable, we were able to predict survival up to 10 years after diagnosis, even for the patients diagnosed most recently (ie, 2010–11). All models were fitted with the STATA command `stpm2` using STATA 13.1.^{17,18}

We included all patients diagnosed during the 40 years from 1971 to 2011 in the models to estimate survival trends, but we report estimates for each cancer survival index at 1, 5, and 10 years after diagnosis only for six selected periods of diagnosis: 1971–72, 1980–81, 1990–91, 2000–01, 2005–06, and 2010–11. We define the difference in net survival between the oldest (75–99 years) and youngest (15–44 years) groups as the age gap in survival. We provide a simple summary of changes in survival by age as the absolute change (%) in the age gap since 1971. A negative value for this change means that the age gap has become wider. For Wales, reliable estimates of net survival could not be obtained for 11.5% of the age-sex-cancer combinations because

	1971-72			1980-81			1990-91			2000-01			2005-06			2010-11 (prediction)		
	1year	5years	10years	1year	5years	10years	1year	5years	10years	1year	5years	10years	1year	5years	10years	1year	5years	10years
All cancers combined																		
All patients	50.1%	29.8%	24.0%	55.8%	35.3%	28.8%	60.6%	41.0%	34.4%	64.9%	47.4%	41.6%	67.6%	50.9%	45.8%	70.5%	54.3%	49.8%
Men	44.7%	25.2%	19.9%	50.6%	29.6%	23.3%	55.7%	34.8%	28.0%	60.7%	42.0%	36.0%	63.7%	45.8%	41.0%	66.7%	49.2%	45.7%
Women	55.5%	34.3%	27.9%	61.0%	40.9%	34.1%	65.3%	47.2%	40.7%	69.0%	52.7%	47.0%	71.5%	56.0%	50.5%	74.2%	59.2%	53.8%
Oesophagus																		
All patients	15.0%	4.3%	3.5%	19.1%	5.3%	4.3%	24.2%	6.5%	5.1%	31.1%	8.8%	7.0%	36.4%	11.5%	9.3%	42.0%	15.3%	12.4%
Men	14.7%	4.0%	3.3%	18.5%	4.8%	3.8%	24.1%	6.1%	4.8%	32.5%	9.1%	7.3%	38.3%	12.0%	9.4%	44.3%	15.6%	12.0%
Women	15.6%	4.8%	3.9%	20.0%	6.2%	5.0%	24.3%	7.1%	5.6%	28.8%	8.2%	6.5%	33.4%	10.8%	9.1%	38.6%	14.7%	13.1%
Stomach																		
All patients	15.4%	5.2%	4.0%	20.6%	8.2%	6.7%	26.8%	10.9%	8.9%	33.9%	14.1%	11.3%	37.8%	16.3%	13.1%	41.7%	18.8%	15.0%
Men	15.3%	5.2%	4.0%	20.7%	8.1%	6.7%	27.0%	10.6%	8.6%	34.7%	13.9%	11.0%	39.3%	16.5%	13.0%	43.8%	19.5%	15.3%
Women	15.5%	5.3%	4.0%	20.5%	8.4%	6.8%	26.5%	11.5%	9.4%	32.4%	14.5%	11.8%	35.2%	16.1%	13.1%	37.9%	17.7%	14.4%
Colon																		
All patients	41.5%	24.6%	22.8%	54.0%	34.2%	31.8%	62.1%	41.6%	38.6%	66.7%	47.5%	44.5%	70.3%	52.6%	50.3%	73.9%	58.2%	56.9%
Men	42.6%	25.3%	23.0%	55.2%	34.6%	31.5%	63.5%	41.9%	38.1%	68.1%	47.6%	43.6%	71.9%	52.9%	49.4%	76.1%	59.2%	56.5%
Women	40.4%	23.8%	22.6%	52.7%	33.8%	32.1%	60.7%	41.3%	39.0%	65.4%	47.5%	45.4%	68.6%	52.3%	51.1%	71.7%	57.3%	57.4%
Rectum																		
All patients	53.3%	24.2%	20.1%	60.6%	32.5%	28.2%	67.8%	42.0%	37.7%	74.0%	51.2%	47.1%	76.7%	55.5%	51.7%	79.2%	59.7%	56.1%
Men	54.1%	23.6%	19.1%	61.4%	32.0%	27.1%	68.7%	41.7%	36.7%	74.8%	51.0%	46.4%	77.5%	55.4%	51.0%	79.9%	59.6%	55.5%
Women	52.2%	25.0%	21.6%	59.5%	33.2%	29.6%	66.6%	42.4%	39.0%	72.8%	51.4%	48.2%	75.6%	55.7%	52.7%	78.1%	59.8%	57.0%
Pancreas																		
All patients	10.6%	2.3%	1.2%	12.1%	2.8%	1.5%	13.0%	2.8%	1.5%	14.7%	2.7%	1.2%	17.4%	3.0%	1.2%	20.9%	3.3%	1.1%
Men	10.2%	2.4%	1.3%	12.4%	3.1%	1.7%	13.5%	3.2%	1.7%	15.3%	3.0%	1.4%	18.1%	3.2%	1.2%	21.7%	3.6%	1.1%
Women	11.0%	2.2%	1.1%	11.9%	2.4%	1.2%	12.5%	2.4%	1.3%	14.0%	2.4%	1.1%	16.7%	2.7%	1.2%	20.2%	3.1%	1.1%
Larynx																		
Men	80.7%	60.2%	50.4%	81.7%	62.1%	52.6%	82.8%	64.1%	54.9%	83.7%	66.0%	57.0%	84.2%	67.0%	58.2%	84.7%	67.9%	59.2%
Lung																		
All patients	16.0%	4.6%	3.1%	18.3%	5.5%	3.7%	20.5%	6.0%	3.8%	24.4%	6.9%	4.0%	28.0%	8.0%	4.4%	32.2%	9.6%	5.0%
Men	16.3%	4.8%	3.2%	18.6%	5.8%	3.9%	20.4%	6.1%	3.9%	23.9%	6.6%	3.7%	27.0%	7.4%	3.8%	30.5%	8.4%	4.0%
Women	15.4%	4.3%	2.9%	17.8%	5.0%	3.2%	20.7%	5.9%	3.7%	25.2%	7.4%	4.5%	29.7%	9.1%	5.4%	35.1%	11.6%	6.6%
Melanoma of skin																		
All patients	81.6%	52.3%	46.4%	88.7%	66.4%	60.4%	93.1%	77.2%	71.9%	95.5%	83.8%	79.7%	96.4%	87.0%	84.4%	97.4%	90.4%	89.8%
Men	74.5%	40.5%	34.9%	84.5%	56.4%	49.8%	90.8%	69.8%	63.4%	94.0%	78.4%	73.3%	95.2%	82.6%	79.3%	96.6%	87.8%	86.8%
Women	86.7%	61.1%	54.9%	91.8%	73.7%	68.3%	94.9%	82.6%	78.2%	96.6%	87.8%	84.5%	97.3%	90.2%	88.3%	97.9%	92.4%	92.1%
Breast																		
Women	81.9%	52.7%	40.1%	85.9%	61.2%	48.4%	89.5%	71.1%	60.0%	92.7%	80.2%	71.6%	94.5%	83.9%	75.6%	96.0%	86.7%	78.5%
Cervix																		
Women	74.0%	51.3%	46.0%	78.6%	58.3%	52.4%	81.6%	62.6%	57.2%	82.8%	65.4%	60.7%	82.6%	66.3%	61.9%	82.9%	67.5%	63.1%
Uterus																		
Women	75.6%	59.0%	55.5%	79.5%	65.1%	61.5%	83.3%	69.5%	65.6%	86.9%	73.1%	69.7%	88.7%	75.9%	73.3%	90.3%	78.8%	77.4%
Ovary																		
Women	43.7%	20.5%	17.9%	50.2%	24.9%	21.5%	57.0%	30.8%	26.4%	64.7%	38.4%	31.7%	68.8%	42.4%	33.5%	72.7%	46.4%	34.8%
Prostate																		
Men	66.1%	36.9%	25.1%	71.5%	38.2%	24.4%	79.6%	49.6%	34.1%	89.5%	73.8%	62.4%	92.4%	81.4%	75.1%	94.0%	84.8%	83.6%
Testis																		
Men	83.3%	70.5%	69.2%	91.2%	84.0%	83.3%	95.8%	92.3%	91.9%	98.0%	96.3%	96.2%	98.7%	97.5%	97.4%	99.1%	98.3%	98.2%
Kidney																		
All patients	44.9%	28.5%	23.0%	51.3%	34.1%	27.6%	57.1%	39.4%	32.3%	62.8%	44.8%	37.9%	67.2%	49.8%	43.0%	72.5%	56.3%	49.6%
Men	45.4%	28.9%	23.0%	52.6%	35.3%	28.5%	58.7%	40.8%	33.4%	63.9%	45.2%	37.8%	68.0%	50.0%	42.9%	73.2%	56.7%	50.0%
Women	43.9%	28.0%	23.1%	49.1%	32.2%	26.1%	54.4%	37.1%	30.5%	60.9%	44.0%	38.0%	65.9%	49.4%	43.2%	71.3%	55.6%	48.9%

(Table 2 continues on next page)

	1971-72			1980-81			1990-91			2000-01			2005-06			2010-11 (prediction)		
	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years
(Continued from previous page)																		
Bladder																		
All patients	60.2%	39.3%	32.4%	73.4%	56.0%	48.0%	77.2%	60.8%	52.8%	74.7%	56.4%	49.5%	73.5%	54.8%	49.2%	72.4%	53.4%	49.5%
Men	62.8%	40.9%	33.7%	76.0%	57.9%	49.3%	80.1%	63.0%	54.2%	78.5%	59.2%	52.0%	77.6%	57.8%	52.4%	76.6%	56.5%	53.5%
Women	53.4%	35.2%	29.0%	66.6%	50.8%	44.7%	69.6%	54.9%	49.0%	64.7%	49.1%	43.0%	63.0%	47.0%	40.9%	61.4%	45.3%	39.1%
Brain																		
All patients	17.7%	7.2%	5.4%	23.3%	9.8%	7.2%	27.7%	11.8%	8.4%	30.4%	12.7%	8.8%	34.7%	15.0%	10.6%	40.1%	18.5%	13.5%
Men	17.6%	6.6%	5.0%	23.3%	9.2%	6.7%	27.9%	11.2%	7.9%	30.9%	12.1%	8.3%	35.3%	14.2%	9.9%	41.1%	17.8%	12.8%
Women	17.9%	7.9%	6.0%	23.3%	10.6%	7.8%	27.4%	12.7%	9.2%	29.8%	13.7%	9.5%	33.9%	16.1%	11.5%	38.8%	19.5%	14.5%
Hodgkin's disease																		
All patients	75.6%	56.5%	47.7%	82.7%	66.8%	58.8%	87.6%	75.1%	69.2%	90.0%	80.3%	75.8%	90.8%	82.9%	78.3%	91.4%	85.0%	80.0%
Men	73.9%	54.2%	45.2%	82.2%	65.1%	56.5%	87.5%	74.6%	68.7%	89.7%	80.4%	75.8%	90.3%	82.5%	77.2%	90.8%	84.1%	77.7%
Women	77.8%	59.4%	51.0%	83.3%	69.2%	61.8%	87.7%	75.8%	69.9%	90.3%	80.2%	75.8%	91.4%	83.4%	79.7%	92.3%	86.3%	83.1%
Non-Hodgkin lymphoma																		
All patients	49.5%	29.9%	22.0%	58.8%	37.5%	28.1%	65.8%	44.9%	35.2%	70.1%	52.3%	43.9%	74.3%	59.7%	52.6%	79.6%	68.8%	63.1%
Men	49.4%	29.3%	21.7%	58.6%	36.8%	27.6%	65.7%	44.2%	34.5%	70.0%	51.6%	43.4%	74.4%	59.1%	51.9%	79.8%	68.1%	62.2%
Women	49.6%	30.6%	22.3%	59.0%	38.4%	28.8%	66.0%	45.8%	35.9%	70.2%	53.2%	44.6%	74.3%	60.5%	53.3%	79.4%	69.5%	64.1%
Multiple myeloma																		
All patients	37.4%	11.8%	6.2%	48.4%	17.2%	8.6%	57.4%	22.0%	10.8%	64.5%	27.7%	14.3%	70.6%	36.0%	21.4%	76.7%	47.0%	32.6%
Men	36.8%	12.1%	6.8%	47.8%	17.2%	9.0%	57.4%	22.2%	11.1%	65.7%	28.8%	15.1%	71.8%	37.9%	23.5%	78.0%	50.0%	36.8%
Women	38.0%	11.4%	5.5%	49.0%	17.1%	8.1%	57.3%	21.8%	10.3%	63.2%	26.4%	13.4%	69.3%	34.0%	19.2%	75.3%	43.8%	27.9%
Leukaemia																		
All patients	34.2%	13.1%	6.9%	47.3%	23.6%	14.9%	57.8%	34.0%	24.0%	63.8%	41.6%	32.3%	66.3%	46.4%	38.7%	68.6%	51.5%	46.1%
Men	35.4%	13.1%	6.6%	48.6%	23.7%	14.4%	59.4%	34.4%	23.6%	65.6%	42.4%	32.3%	68.3%	47.7%	39.4%	70.7%	53.3%	47.6%
Women	32.5%	13.0%	7.2%	45.6%	23.5%	15.6%	55.8%	33.6%	24.6%	61.4%	40.5%	32.2%	63.7%	44.6%	37.8%	65.9%	49.1%	44.2%
Other cancers*																		
All patients	55.3%	38.4%	34.8%	54.7%	36.5%	32.0%	54.5%	35.2%	30.2%	56.6%	37.1%	32.5%	59.7%	40.6%	36.6%	63.5%	45.2%	41.9%
Men	57.3%	40.4%	36.9%	54.3%	35.2%	30.7%	52.6%	31.9%	26.9%	55.0%	33.7%	29.2%	58.7%	37.8%	33.9%	63.1%	43.3%	40.1%
Women	53.0%	36.2%	32.5%	55.2%	37.9%	33.4%	56.6%	39.0%	33.9%	58.4%	41.0%	36.3%	60.9%	43.9%	39.7%	63.9%	47.5%	44.0%

*Other cancers: all other malignant tumours are combined; they also include laryngeal cancer in women and breast cancer in men.

Table 2: 40-year trends in the index of net survival for all cancers combined at 1, 5, and 10 years after diagnosis in adults (15–99 years) in England from 1971 to 2011 and trends in the age-adjusted net survival for 21 selected cancers in England from 1971 to 2011 by sex

of the small number of patients, and broader age groups were constructed to re-estimate survival for those combinations.

Role of the funding source

The funder had no role in study design, quality control, analysis, interpretation of the results, drafting, or the decision to submit for publication. The corresponding author had full access to all data and was responsible for the decision to publish.

Results

The index of net survival for all cancers combined at 1, 5, and 10 years since diagnosis increased substantially between 1971 and 2011 in England and Wales (figure 1, tables 2 and 3). The all-cancers survival index was 50% at 1 year after diagnosis for patients diagnosed in 1971–72. For patients diagnosed during 2005–06, the index was 50% at 5 years after diagnosis, and for patients diagnosed

during 2010–11, we predict that the all-cancers survival index will reach 50% at 10 years after diagnosis.

For patients diagnosed during 2010–11, the survival index for all cancers combined had reached 69–70% at 1 year and a predicted value of 54% at 5 years for both sexes combined. The 5-year survival index rose by 24% (from 30% to 54%) and the 10-year survival index by 26% (from 24% to 50%) between the periods 1971–72 and 2010–11. Most of the increase occurred between 1990 and 2011.

The survival index for all cancers combined is on average 10% higher for women than for men at each time interval since diagnosis. The pattern of increase in the index was fairly similar for both men and women during the whole period, although the increase was linear for women but it became steeper for men after 1990–91. For patients diagnosed during 2010–11, the all-cancers survival index for women in England was 74% at 1 year, 59% at 5 years, and 54% at 10 years, whereas the figures for men were 67% at 1 year, 49% at 5 years, and

	1971-72			1980-81			1990-91			2000-01			2005-06			2010-11 (prediction)		
	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years
All cancers combined																		
All patients	48.1%	28.9%	23.4%	53.6%	34.7%	28.9%	58.4%	40.4%	34.6%	63.2%	46.9%	41.6%	66.3%	50.6%	46.0%	69.4%	54.2%	50.2%
Men	42.9%	24.8%	20.4%	48.2%	28.8%	23.7%	53.2%	33.9%	28.1%	59.1%	41.5%	35.9%	62.7%	45.7%	41.0%	65.9%	49.2%	45.5%
Women	53.2%	32.9%	26.3%	58.9%	40.5%	34.1%	63.4%	46.8%	41.1%	67.2%	52.2%	47.2%	69.9%	55.5%	51.0%	72.8%	59.0%	54.8%
Oesophagus																		
All patients	16.9%	5.2%	4.1%	18.7%	6.0%	5.2%	22.8%	6.9%	5.8%	30.7%	8.8%	6.7%	35.5%	10.6%	7.9%	39.7%	12.9%	9.5%
Men	17.9%	5.1%	3.8%	19.1%	5.8%	4.9%	23.2%	7.0%	6.0%	32.8%	9.3%	7.1%	37.7%	10.9%	7.8%	42.3%	12.7%	8.7%
Women	15.2%	5.4%	4.8%	18.1%	6.4%	5.6%	22.1%	6.8%	5.5%	27.4%	7.9%	6.1%	32.1%	10.3%	8.0%	35.8%	13.3%	10.8%
Stomach																		
All patients	15.2%	5.7%	4.6%	21.3%	10.1%	8.9%	24.7%	10.8%	9.2%	30.9%	12.6%	9.9%	36.5%	15.5%	12.0%	43.1%	19.5%	14.9%
Men	15.3%	5.6%	4.5%	21.0%	9.7%	8.6%	25.0%	10.6%	9.1%	32.3%	12.6%	9.9%	38.2%	15.5%	11.7%	45.0%	19.4%	14.4%
Women	15.0%	6.0%	4.9%	21.9%	10.8%	9.4%	24.1%	11.2%	9.3%	28.5%	12.7%	10.1%	33.5%	15.6%	12.4%	39.6%	19.6%	16.0%
Colon																		
All patients	42.7%	25.0%	22.8%	51.8%	33.3%	30.9%	58.4%	39.8%	37.2%	63.2%	45.2%	42.4%	67.8%	50.9%	48.3%	73.0%	57.7%	55.4%
Men	43.1%	26.5%	24.5%	51.9%	33.3%	30.9%	60.0%	40.3%	37.4%	65.8%	46.6%	43.3%	70.2%	51.8%	48.5%	74.9%	57.9%	54.9%
Women	42.2%	23.4%	21.2%	51.8%	33.3%	31.0%	56.9%	39.2%	37.0%	60.5%	43.7%	41.6%	65.4%	49.9%	48.0%	71.1%	57.5%	55.8%
Rectum																		
All patients	50.8%	22.9%	19.7%	58.5%	31.2%	27.7%	65.7%	40.6%	37.1%	72.4%	50.0%	46.7%	75.2%	54.4%	51.3%	77.7%	58.5%	55.6%
Men	50.6%	21.4%	17.9%	58.7%	29.9%	26.1%	66.5%	39.8%	35.9%	73.2%	49.5%	45.8%	76.1%	54.1%	50.6%	78.6%	58.4%	55.1%
Women	51.0%	25.2%	22.1%	58.1%	33.1%	30.0%	64.6%	41.7%	38.9%	71.4%	50.8%	48.0%	74.0%	54.8%	52.3%	76.4%	58.6%	56.4%
Pancreas																		
All patients	12.2%	3.8%	2.4%	12.8%	4.6%	3.4%	12.9%	4.2%	2.8%	14.0%	3.0%	1.5%	16.3%	3.0%	1.3%	19.0%	3.3%	1.2%
Men	11.5%	4.0%	2.7%	13.0%	5.6%	4.6%	13.5%	5.0%	3.7%	14.8%	3.4%	1.8%	16.7%	3.4%	1.5%	19.4%	3.7%	1.4%
Women	12.9%	3.7%	2.1%	12.5%	3.7%	2.3%	12.4%	3.3%	2.0%	13.3%	2.6%	1.3%	15.8%	2.7%	1.2%	18.6%	2.9%	1.1%
Larynx																		
Men	77.7%	56.3%	45.9%	82.5%	64.8%	55.6%	82.1%	63.9%	54.5%	80.2%	60.4%	50.4%	81.4%	63.3%	53.7%	84.0%	68.1%	59.5%
Lung																		
All patients	15.6%	5.1%	3.6%	18.7%	7.2%	5.5%	19.7%	6.8%	4.7%	21.5%	5.9%	3.3%	25.5%	6.9%	3.6%	31.1%	8.6%	4.2%
Men	14.6%	4.2%	2.8%	18.6%	7.2%	5.6%	19.5%	6.7%	4.6%	21.1%	5.5%	2.9%	24.4%	6.3%	3.1%	28.8%	7.7%	3.7%
Women	17.4%	6.6%	5.1%	18.8%	7.0%	5.3%	20.1%	6.9%	4.9%	22.2%	6.6%	4.0%	27.4%	8.0%	4.3%	35.2%	10.3%	5.1%
Melanoma of skin																		
All patients	79.9%	51.1%	44.0%	82.3%	63.1%	57.2%	85.6%	71.4%	66.3%	91.3%	77.5%	72.9%	94.4%	82.4%	77.6%	96.8%	89.0%	82.1%
Men	73.8%	38.9%	33.3%	76.6%	51.0%	44.6%	81.8%	62.5%	55.9%	89.4%	71.0%	65.8%	93.1%	76.4%	68.9%	95.8%	83.7%	68.3%
Women	84.4%	60.1%	52.0%	86.5%	72.0%	66.6%	88.3%	78.0%	73.9%	92.7%	82.2%	78.1%	95.3%	86.7%	84.1%	97.6%	92.9%	92.2%
Breast																		
Women	74.9%	47.9%	34.8%	81.8%	60.3%	48.5%	87.4%	71.7%	62.3%	91.4%	80.4%	73.4%	93.0%	83.8%	77.9%	94.3%	86.7%	81.8%
Cervix																		
Women	73.9%	52.8%	47.4%	80.0%	63.2%	57.8%	78.6%	59.9%	55.0%	78.5%	59.9%	55.2%	79.7%	62.4%	57.5%	81.7%	65.6%	60.3%
Uterus																		
Women	72.7%	55.9%	53.4%	76.2%	61.7%	56.8%	80.6%	67.0%	62.2%	85.3%	72.4%	69.6%	88.1%	76.8%	73.9%	90.5%	81.2%	77.8%
Ovary																		
Women	48.2%	22.2%	18.0%	52.0%	26.2%	21.8%	56.9%	31.4%	26.6%	61.1%	36.6%	31.8%	63.1%	39.2%	34.4%	65.1%	41.9%	37.1%
Prostate																		
Men	62.7%	36.6%	27.8%	65.9%	35.9%	25.6%	72.9%	44.6%	32.9%	85.0%	68.8%	59.1%	90.1%	79.8%	74.9%	93.7%	87.1%	87.1%
Testis																		
Men	82.9%	69.5%	66.2%	89.9%	81.1%	80.0%	94.4%	89.7%	89.1%	97.1%	95.0%	94.1%	97.4%	96.0%	94.4%	97.4%	96.6%	93.9%
Kidney																		
All patients	43.7%	29.0%	24.4%	46.9%	31.0%	25.1%	53.0%	36.5%	29.7%	61.6%	46.2%	39.5%	66.6%	51.2%	44.0%	70.8%	55.2%	47.3%
Men	44.8%	30.6%	25.3%	48.5%	32.0%	25.6%	54.6%	37.2%	30.0%	62.3%	46.9%	39.9%	67.6%	51.4%	43.5%	72.2%	53.9%	44.2%
Women	41.9%	26.4%	22.9%	44.2%	29.5%	24.4%	50.3%	35.4%	29.2%	60.5%	45.1%	38.7%	64.8%	50.8%	44.8%	68.5%	57.3%	52.4%

(Table 3 continues on next page)

	1971-72			1980-81			1990-91			2000-01			2005-06			2010-11 (prediction)		
	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years	1 year	5 years	10 years
(Continued from previous page)																		
Bladder																		
All patients	53.8%	37.4%	33.9%	66.3%	49.1%	42.5%	77.1%	61.5%	53.5%	81.4%	67.6%	61.6%	78.0%	63.7%	60.6%	70.5%	55.5%	56.8%
Men	56.1%	38.0%	34.1%	69.5%	51.5%	44.6%	80.3%	64.6%	56.3%	85.0%	71.0%	64.8%	81.8%	66.7%	63.7%	74.5%	57.9%	59.9%
Women	47.9%	35.8%	33.4%	58.0%	43.1%	37.3%	68.9%	53.7%	46.2%	72.2%	58.9%	53.6%	68.3%	56.0%	52.8%	60.1%	49.1%	48.8%
Brain																		
All patients	24.4%	10.7%	7.9%	26.7%	11.8%	8.9%	29.0%	12.8%	9.6%	33.1%	14.8%	10.6%	36.8%	16.5%	11.4%	40.1%	18.0%	12.0%
Men	24.5%	10.3%	7.7%	26.6%	11.5%	8.7%	28.3%	11.9%	8.8%	32.7%	13.5%	9.1%	36.9%	15.6%	10.3%	40.8%	17.5%	11.2%
Women	24.4%	11.1%	8.2%	26.7%	12.3%	9.1%	29.9%	14.2%	10.8%	33.7%	16.6%	12.7%	36.6%	17.7%	13.0%	39.2%	18.6%	13.1%
Hodgkin's disease																		
All patients	72.1%	52.1%	43.1%	78.2%	62.0%	54.0%	84.4%	72.0%	65.4%	87.6%	78.5%	73.2%	89.7%	81.5%	76.8%	92.3%	85.7%	81.8%
Men	74.5%	54.8%	44.5%	79.1%	62.9%	53.7%	85.1%	73.0%	65.6%	87.4%	78.3%	72.2%	89.6%	81.6%	76.2%	92.3%	85.6%	81.0%
Women	68.9%	48.6%	41.2%	77.0%	60.9%	54.4%	83.5%	70.7%	65.2%	87.9%	78.7%	74.6%	89.8%	81.4%	77.7%	92.4%	85.8%	82.8%
Non-Hodgkin lymphoma																		
All patients	50.2%	31.1%	23.8%	54.7%	33.8%	25.4%	61.0%	39.8%	30.8%	68.6%	50.7%	41.9%	73.7%	58.3%	50.3%	79.3%	66.7%	59.7%
Men	51.8%	30.8%	22.1%	54.2%	33.4%	24.2%	60.0%	39.7%	30.0%	68.9%	50.9%	41.3%	74.0%	57.9%	48.7%	79.0%	65.1%	56.8%
Women	48.3%	31.5%	25.9%	55.3%	34.2%	26.7%	62.2%	40.0%	31.7%	68.4%	50.4%	42.7%	73.4%	58.8%	52.0%	79.6%	68.5%	63.1%
Multiple myeloma																		
All patients	34.1%	12.6%	8.0%	49.1%	19.9%	11.9%	57.8%	24.1%	13.5%	62.7%	26.9%	14.0%	67.6%	33.6%	19.0%	73.8%	44.5%	28.7%
Men	33.2%	14.0%	10.7%	48.6%	20.0%	12.7%	58.3%	24.3%	13.8%	64.8%	28.8%	15.9%	70.0%	35.8%	20.9%	76.2%	46.7%	30.2%
Women	35.2%	11.1%	5.1%	49.7%	19.8%	11.0%	57.2%	23.9%	13.1%	60.4%	24.7%	11.9%	64.9%	31.2%	16.8%	71.0%	42.0%	27.0%
Leukaemia																		
All patients	30.2%	11.0%	6.1%	43.5%	21.2%	14.1%	55.4%	33.0%	24.5%	64.9%	43.6%	34.1%	69.1%	49.5%	40.5%	72.9%	55.6%	47.7%
Men	27.7%	8.7	3.9%	43.5%	20.2%	12.8%	57.0%	33.1%	24.0%	66.5%	43.3%	32.5%	70.4%	49.4%	39.4%	74.3%	56.2%	47.9%
Women	33.4%	14.0%	8.9%	43.5%	22.4%	15.6%	53.4%	32.7%	25.1%	62.9%	44.0%	36.2%	67.5%	49.7%	42.0%	71.0%	54.7%	47.4%
Other cancers*																		
All patients	53.9%	37.6%	33.7%	55.7%	39.3%	34.9%	55.8%	38.9%	34.1%	55.9%	38.3%	33.4%	58.9%	41.1%	36.1%	62.9%	45.2%	40.3%
Men	56.4%	40.2%	36.3%	56.4%	39.8%	35.1%	55.2%	37.5%	32.6%	54.6%	35.5%	30.7%	58.5%	38.9%	33.8%	64.1%	44.3%	38.6%
Women	51.1%	34.7%	30.7%	54.8%	38.7%	34.6%	56.5%	40.5%	35.9%	57.4%	41.4%	36.4%	59.3%	43.6%	38.9%	61.5%	46.4%	42.2%

*Other cancers: all other malignant tumours are combined; they also include laryngeal cancer in women and breast cancer in men

Table 3: 40-year trends in the index of net survival for all cancers combined at 1, 5, and 10 years after diagnosis in adults (15–99 years) in Wales from 1971 to 2011 and trends in the age-adjusted net survival for 21 selected cancers in Wales from 1971 to 2011 by sex

46% at 10 years. Both the levels and the trends in the all-cancers survival index were similar in England and Wales. The average absolute difference between the two countries was less than 1% (figure 1, tables 2 and 3).

Survival for both sexes combined varied widely for different cancers, with the most recent predicted 10-year net survival adjusted for age and sex ranging from only 1.1% for pancreatic cancer to 98.2% for testicular cancer. A scatter-plot of the 1-year, 5-year, and 10-year survival estimates for adults diagnosed in 2010–11 against the absolute change since 1971 enables three broad clusters of cancers to be identified (figure 2). The first cluster consists of cancers with high survival in 2010–11 for which the absolute increase in survival since 1971–72 is progressively larger for survival at 1, 5, and 10 years. It includes cancers of the breast, prostate, testis, and uterus, and melanoma and Hodgkin's disease.

The second cluster is of cancers with a moderate level of survival (64–84%) in 2010–11 and, generally, smaller

increases since 1971–72. This cluster consists of cancers of the larynx, cervix, rectum, colon, bladder, ovary, and kidney, with non-Hodgkin lymphoma, multiple myeloma, and leukaemia. For multiple myeloma and leukaemia, age-adjusted 10-year survival rose by more than 22% between the periods 1990–91 and 2010–11, from around 10.8% to a predicted 32.6% for multiple myeloma and from 24.0% to 46.1% for leukaemia (table 2).

The third cluster is of cancers for which survival for patients diagnosed during 2010–11 is still low, and for which little or no improvement has occurred in the past 40 years: this group consists of malignancies of the brain, stomach, lung, oesophagus, and pancreas.

This clustering can be seen as early as 1 year after diagnosis, and each cancer is in the same cluster, irrespective of the time since diagnosis (and the nation). We observed the largest absolute change in the age-adjusted survival for multiple myeloma, leukaemia, and prostate cancer.

1-year survival from lung cancer has improved substantially, from 16% in 1971–72 to 32% in 2010–11. However, estimated long-term survival for patients diagnosed in 2010–11 is very poor for both sexes: as low as 10% at 5 years and 4% and 7% in men and women, respectively, at 10 years. This overall pattern of no improvement in long-term survival is common in the cluster of poor-prognosis cancers (oesophagus, stomach, pancreas, and brain), for men and women and for both England and Wales.

Survival for breast cancer has seen a rapid and substantial improvement during the past 40 years. 5-year survival increased from 53% in 1971–72 to a predicted value of 87% in 2010–11. After 10 years, survival rose from 40% in 1971–72 to a predicted 78% for patients diagnosed during 2010–11. Differences between 5-year and 10-year survival estimates remained broadly constant since 1971, showing that most of the improvements in long-term survival arose in the first 5 years after diagnosis. Breast cancer accounted for nearly a third of all cancers in women, which partly explains the higher all-cancers survival index in women than in men.

Although survival from cancers of the colon and rectum is much lower than survival from breast cancer (around 20% lower in 2010–11), the trends in 1-year, 5-year, and 10-year survival for these two cancers have followed an almost identical pattern to that of breast cancer during the past 40 years.

For men diagnosed with prostate cancer during 2010–11, the predicted values for 5-year and 10-year estimates are almost identical at 85% and 84%, respectively, which are huge increases from the values of 37% and 25% for men diagnosed 40 years ago. The trends are quite distinct for short-term, medium-term, and long-term survival. In both England and Wales, 1-year survival has been increasing since 1971–72, whereas acceleration in 5-year survival started for men diagnosed in the 1980s; 10-year survival only began increasing for men diagnosed in the 1990s.

For women diagnosed with cancer of the ovary during 2010–11, the age-adjusted survival was predicted as 46% at 5 years and 35% at 10 years compared with 20% and 18%, respectively, for women diagnosed during 1971–72. These results suggest that the underlying increase in survival of up to 5 years is likely to continue.

Net survival is generally lower for the oldest patients (75–99 years) than the youngest (15–44 years), even though net survival accounts for a higher mortality from causes other than cancer in elderly patients. This finding is shown by a scatter-plot of the age gap in net survival at 1, 5, and 10 years after diagnosis for adults diagnosed in

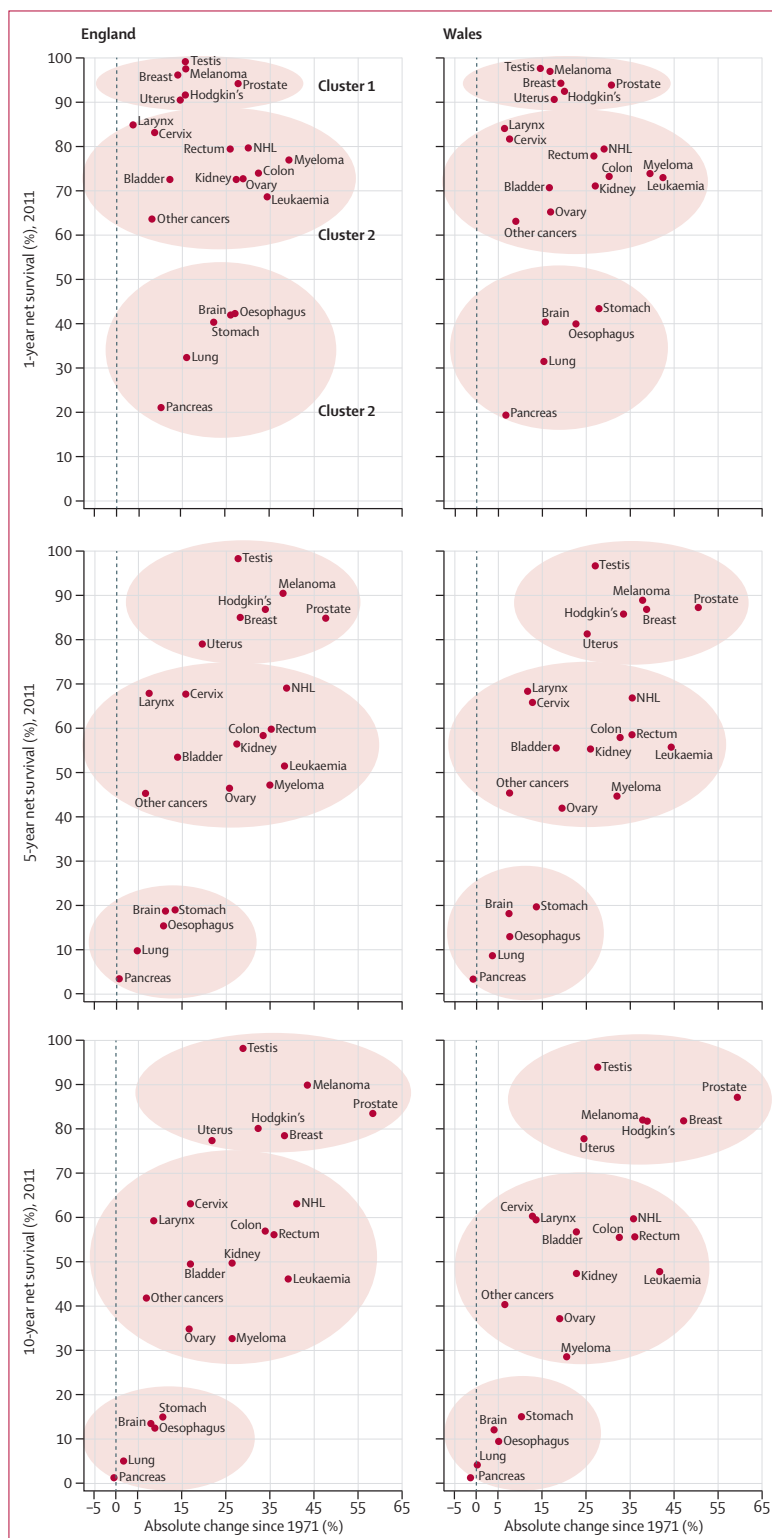
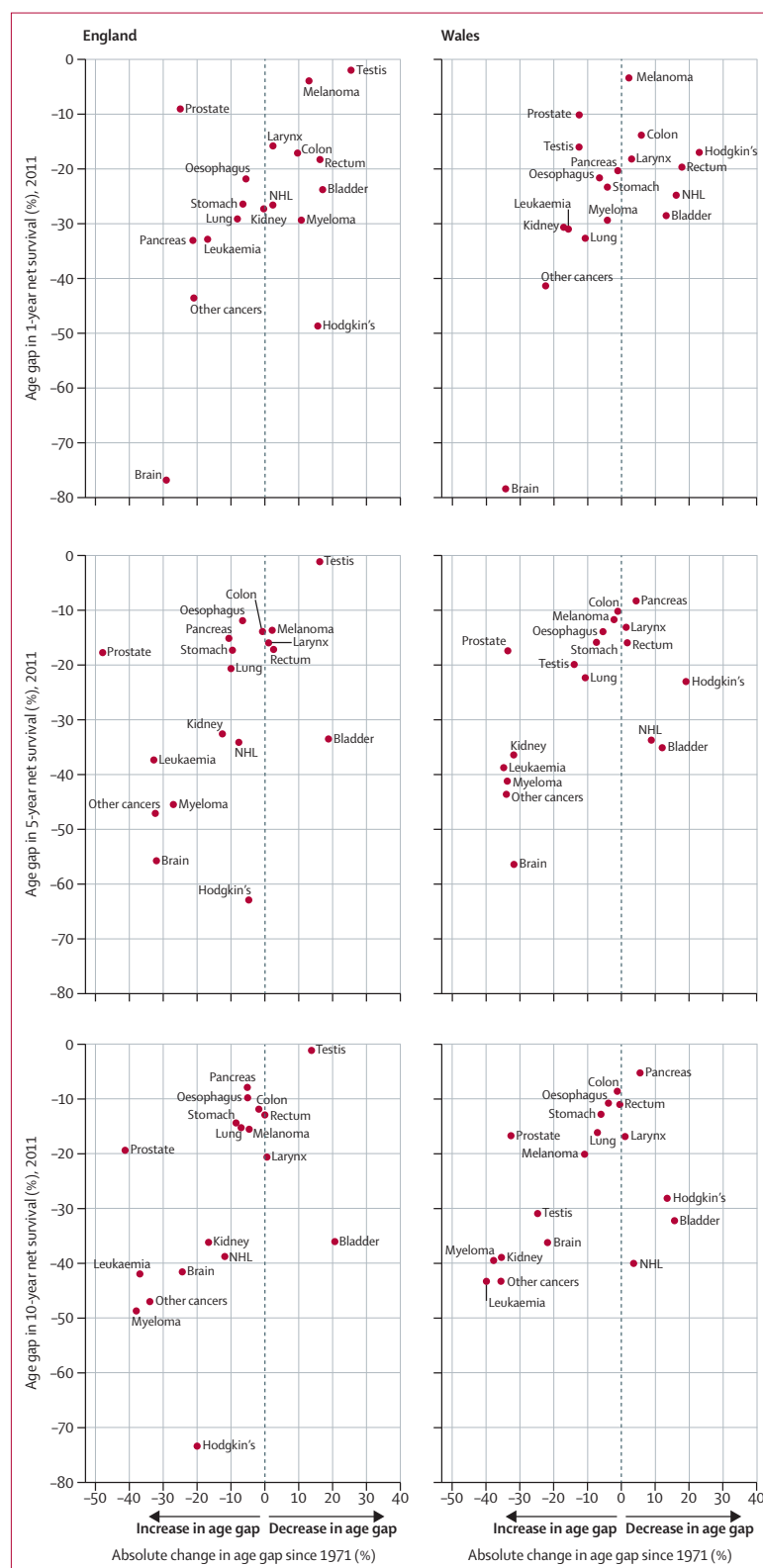


Figure 2: Net survival adjusted for age and sex for each cancer in 2010–11, and absolute change* since 1971, all adults (15–99 years), England and Wales: 1, 5, and 10 years after diagnosis

*The absolute change is the simple arithmetic difference between net survival in 2010–11 and the survival in 1971–72. NHL=non-Hodgkin lymphoma.



2010–11 against the absolute change since 1971–72: it shows a negative gap in survival for most cancers (y-axis of figures 3 and 4).

The largest age gaps in survival in men were observed for cancers for which high-dose chemotherapy is the key treatment (lymphoma, multiple myeloma, and leukaemia), but we could not identify any overall temporal patterns. For women, the largest age gaps were noted for brain tumours, and cancers of the ovary and cervix, and multiple myeloma, but the clustering was less obvious than in men. The age gap tended to narrow for melanoma and cancer of the uterus in women but widened for long-term survival of ovarian cancer.

Discussion

The index of net survival for all cancers combined has increased substantially: for patients diagnosed in 1971–72, the index was 50% at 1 year after diagnosis. Our prediction is that, for patients diagnosed during 2010–11, the all-cancers survival index will reach 50% at 10 years after diagnosis. Very similar patterns of change and levels of survival were noted in both England and Wales.

Survival has increased steadily during the 40 years since 1971, with a slight acceleration in the past 10–15 years, particularly for 5-year and 10-year survival, in both England and Wales. After implementation of the NHS cancer plan for England,¹⁹ we reported a slight acceleration in the 1-year cancer survival trends during 2004–06, by contrast with Wales,² where a national cancer plan was only introduced in 2006. The pattern was not so clear for survival at 3 years after diagnosis. The findings reported here suggest a continuing acceleration of these trends for longer-term survival between 2005–06 and 2010–11 in England, but also in Wales (panel).

The completeness and quality of cancer registration and follow-up data in both England and Wales have been systematically assessed and are thought to be very high throughout the period 1971–2011, despite undeniable improvement during the 1970s–80s.^{21–23} This improvement cannot explain long-term trends in cancer survival.^{24,25} Furthermore, with the exception of bladder cancer, overall changes in disease definitions are limited, even for haemopoietic malignancies. To affect the survival index, such a change in disease definition would need to affect a substantial proportion of all cancers, for which prognosis would also need to be very different from that for other cancers. These conditions are not met.

In some strata defined by age, sex, cancer, and calendar period of diagnosis, especially in Wales, few deaths

Figure 3: Age gap* in net survival by cancer, men (15–99 years) diagnosed during 2010–11 versus absolute change† in the age gap since 1971, England and Wales: 1, 5, and 10 years after diagnosis

*The age gap represents the absolute difference (%) between net survival in the oldest (75–99 years) and youngest (15–45 years) groups of patients; a negative value means that survival is lower in the oldest group than the youngest group.

†The absolute change is the simple arithmetic difference between the age gap in 2010–11 and the age gap in 1971–72. NHL=non-Hodgkin lymphoma.

occurred. To obtain more stable net survival estimates, we therefore estimated net survival using a modelling approach rather than the non-parametric Pohar-Perme approach.²⁰

The index of net survival for all cancers combined provides one convenient number that summarises the overall patterns of cancer survival in any one population or country, in each calendar period for young and old men and women and for a wide range of cancers with very disparate survival. The index is unaffected by changes in the proportion of cancers of different lethality in either sex, such as the reduction of lung cancer or the increase in prostate cancer in men. Similarly, the index is unaffected by ageing of the population of patients with cancer or shifts in the proportion of any cancer between men and women. The value of the index changes only when survival for one or more cancers changes, for one or more age groups. The index therefore shows overall progress in cancer management, whether from earlier diagnosis, or earlier stage of disease, or improved treatment and care.

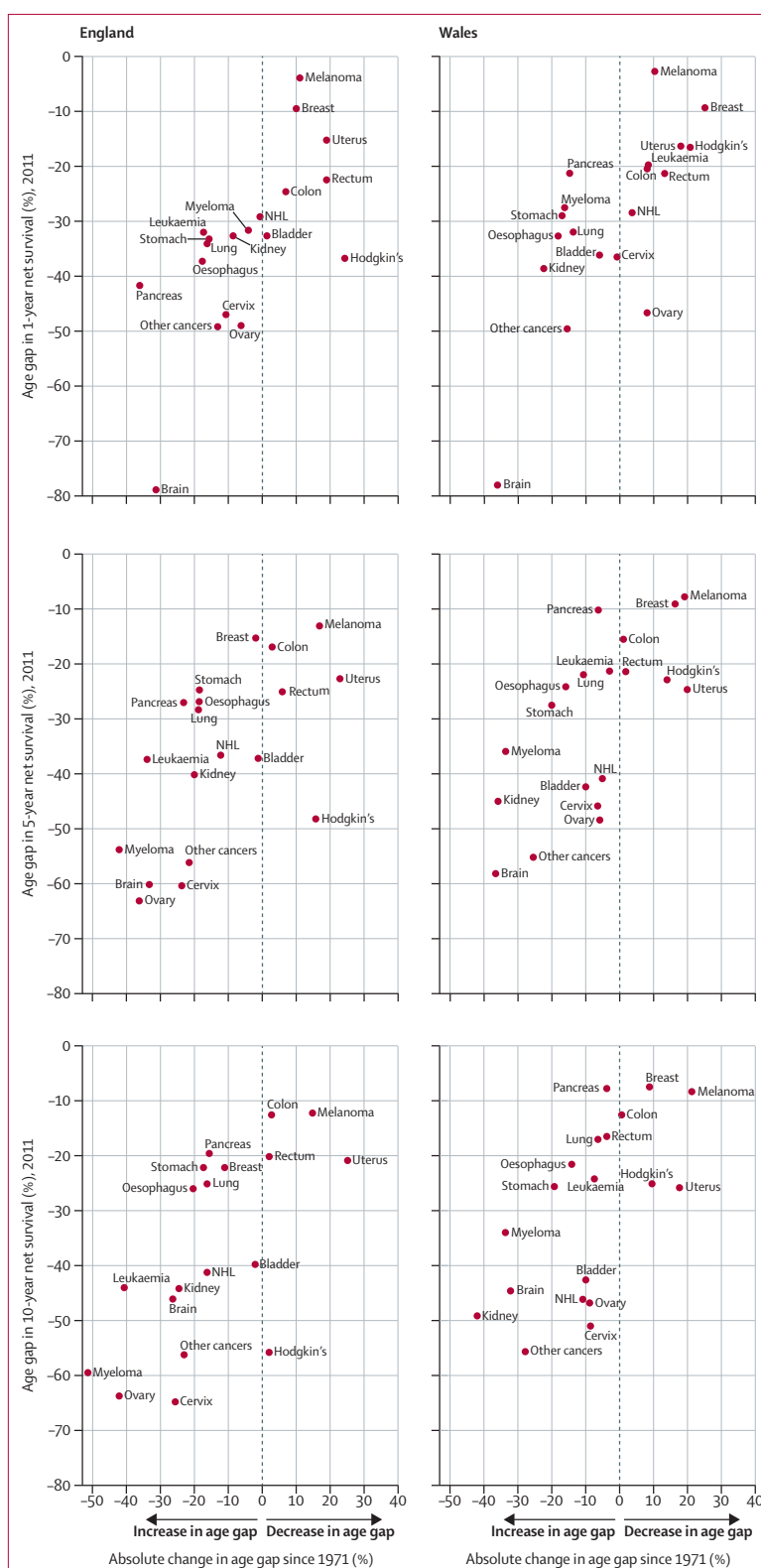
However, the all-cancers survival index needs careful interpretation: for example, the predicted value of 50% for the 10-year all-cancers survival index for 2010–11 does not mean that half of all patients will be cured or “beat cancer”, as has been portrayed in the media.²⁶ The index is designed as a public health measure that summarises cancer survival trends in an entire population, to help to assess progress in the overall effectiveness of the health system in diagnosis and management of patients with cancer. The index does not reflect the prospects of survival for any individual patients with cancer. The index is based on net survival, which is an unbiased measure of population-based survival from cancer after adjustment for other causes of death. Net survival is the most valid available metric for comparison of survival between populations and for assessment of progress in cancer survival over time. The all-cancers net survival index should nevertheless be interpreted in conjunction with other information available in the population or country for which the index has been prepared. It should be seen as a guide to raise questions about the potential for improvement.

The average 10% difference in the survival index between men and women has been a consistent feature for 40 years. It arises because, for several individual cancers, survival is slightly higher for women, but mostly because the cancers that are most common in women, such as breast cancer (weight of 0·31 in the survival index for women), generally have higher survival than the cancers that are most common in men, such as lung

Figure 4: Age gap* in net survival by cancer, women (15–99 years) diagnosed during 2010–11 versus absolute change† (%) in the age gap since 1971, England and Wales: 1, 5, and 10 years after diagnosis

*The age gap represents the absolute difference (%) between net survival in the oldest (75–99 years) and youngest (15–45 years) groups of patients; a negative value means that survival is lower in the oldest group than the youngest group.

†The absolute change is the simple arithmetic difference between the age gap in 2010–11 and the age gap in 1971–72. NHL=non-Hodgkin lymphoma.



Panel: Research in context**Systematic review**

Health policy measures to improve the organisation and delivery of services for the prevention, diagnosis, and treatment of cancer should be based on sound evidence. Population-based survival trends have proved to be a key metric for the overall effectiveness of health systems. An unbiased estimator of net survival was introduced in 2012.²⁰ We have not undertaken a literature review, but so far, only a few countries have published population-based cancer survival using this estimator, including in England by our research group.¹² No other country has constructed a single, summary index of net survival for all cancers combined. A simple, robust, one-number index of net survival for all cancers combined can contribute to the evidence base for rational health policy.

Interpretation

Changes in the net survival index reflect changes in survival for one or more cancers, not simply changes in the distribution of cancer patients by age, cancer site, or sex. The net survival index increased substantially between 1971 and 2011, representing a substantial gain in overall survival from all cancers combined. Net survival varied widely for different cancers, and was generally lower for older patients than younger patients, even after adjustment for the higher mortality from other causes in older patients. Three clusters of cancers, with high, moderate, and low survival, can be distinguished as early as 1 year after diagnosis. Overall, the survival trends are encouraging in both England and Wales, but they also suggest strongly the need for renewed efforts to achieve better outcomes.

See Online for appendix

cancer (weight of 0.22 in the index for men). The slight narrowing in the sex gap observed in the most recent periods might be explained by the rapid increase in survival for prostate cancer (weight of 0.19 in the index for men), particularly at 5 and 10 years after diagnosis. This rapid increase in survival for prostate cancer has been largely attributed to the widespread use of prostate-specific antigen (PSA) testing, resulting in the diagnosis of many less advanced tumours with a shift of the stage distribution to less advanced and less aggressive disease. However, importantly, survival had already started to increase, albeit more slowly, much before PSA testing was widely used.²⁷ The more recent increase in long-term survival suggests that this improvement is not simply because of a shift in the stage distribution after increasingly wide use of the PSA test. The increase in short-term survival, which began as early as the 1970s, and the increase in 5-year survival in the 1980s and then in the 10-year survival in the following decade cannot simply be attributed to PSA.

We were able to group the 21 most common cancers into three clusters on the basis of their survival. Despite some large gains in survival, these clusters are, with few exceptions, the same in 2011 as in 1971 (data not shown).

The clusters are identifiable as early as 1 year after diagnosis, and they are consistent at 5 and 10 years after diagnosis, both in England and Wales.

Cluster 1 includes cancers with a good prognosis: survival is now very high, after a large increase since 1971, particularly at 5 and 10 years after diagnosis. 1-year survival seems to have reached a ceiling for most of these cancers, but survival at 5 and 10 years is still much lower than at 1 year for breast cancer and Hodgkin's disease. The absence of any plateau in survival, even 10 years after diagnosis, shows that cure at the population level has still not been reached for these cancers, leaving room for substantial further improvement in long-term survival.

For most cancers in the other two clusters, survival at 5 and 10 years after diagnosis is still much lower than 1-year survival. The second cluster consists of a further mix of cancers for which either survival has remained moderate since the early 1970s, or moderate levels of survival in 2011 are the result of large improvements during the past 40 years. The second situation is well illustrated by the steep increase in survival from multiple myeloma since 2000–01, probably explained by the introduction of higher-dose treatment regimens around 2000. For the cancers in this cluster that have shown no evidence of improvement, efforts should be made to achieve earlier diagnosis, and to focus on stricter guidelines for improved treatment, such as increased use of surgery, radiotherapy with curative intent, neoadjuvant therapies, or a combination of the three.

The effect of mass-screening on survival varies with the cancer. For cervical cancer, an efficient screening programme does not necessarily lead to an improvement in survival because screening prevents the occurrence of invasive tumours, thereby reducing incidence, and the remaining patients are, on average, diagnosed with more advanced disease.²⁸ A quasi-plateau in 1-year survival has been observed since 2000–01 (appendix 1 and 2).

By contrast, breast cancer screening aims to diagnose the disease at an early stage, rather than to prevent it. Its real effect on survival has been questioned mainly because of possible overdiagnosis and lead time. However, overdiagnosis does not exceed a few percent,²⁹ and the advantage in survival remains important for screen-detected breast cancer after accounting for lead time.³⁰ Improvement in breast cancer survival has been large because of both early diagnosis and improved treatment, although net survival continues to decrease even 10 years after diagnosis, showing late recurrences. The age gap in survival has also decreased, supporting more rapid improvement in survival for older women (and for the screened age group) than in younger women.³¹

Screening for colorectal cancer, which started in 2006, aims to prevent invasive malignant tumours (by removing polyps with adenomatous change) and to diagnose cancer at an early stage. Therefore, although it is too recent to have any effect on these results, lessons from both cervical and breast cancer screening

programmes will also help us to monitor the effect of screening on the prognosis of colorectal cancer.

A wide age gap in survival was still present for most cancers in 2010–11. Some of these differences are related to screening or early diagnostic practices (breast, cervix, prostate). Also, the disease, and its prognosis, might radically differ by age, such as leukaemia: the treatment of acute disease in young patients improved substantially, by contrast with chronic leukaemia in elderly patients, but separation of both diseases is not possible over the entire period 1971–2011. However, in other countries, the age gap in cancer survival is much narrower than in England and Wales.^{32,33} The wide age-related inequalities in cancer survival in England and Wales are thus likely to be avoidable. They could be substantially reduced.

1-year survival has improved substantially for cancers with a particularly poor prognosis (cluster 3), but longer-term survival (5 and 10 years after diagnosis) has hardly changed during the past four decades. Among these cancers, substantial improvements should be achievable for lung cancer: in 2011, National Institute for Health and Care Excellence (NICE) guidelines³⁴ underlined the need for improved staging and increased widespread access to surgery and radiotherapy with curative intent for non-small-cell lung cancer. Adherence to these guidelines and their effect on cancer outcomes has not yet been exhaustively assessed.³⁵

In summary, despite impressive overall improvements in cancer survival during the past 40 years in both England and Wales, the wide and persistent differences in survival between cancers, together with the wide and persistent age gap in survival for most cancers, suggest the need for renewed efforts to achieve improved outcomes, particularly in elderly patients. The findings reported here offer clues for focused research to dissect the underlying causes of these differences in cancer survival. The results should prompt action to improve public health in both England and Wales. This research will need systematic linkage of clinical audit streams and other detailed data streams to population-based cancer registry data, but the recent crisis of public concern about the sharing of individual health data for confidential public health research will need to be resolved first.³⁶

Contributors

MQ did the analysis. MQ and BR designed the analytic strategies and constructed the indexes. MQ, MPC, and BR wrote the Article and interpreted the findings.

Declaration of interests

We declare no competing interests.

Acknowledgments

We thank regional cancer registries in England and Wales for their continuing efforts in collecting data for all cancer patients to the highest standards of quality and completeness and the reviewers for their constructive comments. MQ, MPC, and BR report grants from Cancer Research UK during this study. Grant numbers C1336/A11700 and C7923/A18348.

References

- 1 Coleman MP. Cancer survival: global surveillance will stimulate health policy and improve equity. *Lancet* 2014; **383**: 564–73.
- 2 Cancer Services Co-ordinating Group. Designed to tackle cancer in Wales: a Welsh assembly government policy statement: Welsh assembly government, 2006. <https://www.wales.nhs.uk/documents/Designed-to-tackle-Cancer.pdf> (accessed Nov 20, 2014).
- 3 Expert Advisory Group on Cancer. A policy framework for commissioning cancer services (Calman-Hine report). London: Department of Health, 1995.
- 4 Department of Health. Challenging cancer. London: Department of Health, 1999.
- 5 Department of Health. Cancer Reform Strategy. London: Department of Health, 2007.
- 6 Rachet B, Maringe C, Nur U, et al. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncol* 2009; **10**: 351–69.
- 7 Rachet B, Ellis L, Maringe C, et al. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* 2010; **103**: 446–53.
- 8 Department of Health. Improving outcomes: a strategy for cancer—second annual report 2012. London: Department of Health, 2012.
- 9 Richards MA. The size of the prize for earlier diagnosis of cancer in England. *Br J Cancer* 2009; **101** (suppl 2): S125–29.
- 10 Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009; **101** (suppl 2): S1–4.
- 11 The National Awareness and Early Diagnosis Initiative (NAEDI). NAEDI Newsletter 2009. http://www.cancerresearchuk.org/prod_consump/groups/cr_common/@nre/@hea/documents/generalcontent/013776 (accessed Nov 20, 2014).
- 12 Solomon T, Rachet B, Whitehead S, Coleman MP. Cancer survival in England: patients diagnosed 2007–2011 and followed up to 2012. 2013. <http://www.ons.gov.uk/ons/rel/cancer-unit/cancer-survival/cancer-survival-in-england-patients-diagnosed-2007-2011-and-followed-up-to-2012/stb-cancer-survival-in-england-patients-diagnosed-2007-2011-and-followed-up-to-2012.html> (accessed Nov 20, 2014).
- 13 Cancer Research UK Cancer Survival Group. Life tables for England and Wales by sex, calendar period, region and deprivation. 2009. <http://www.lshtm.ac.uk/eph/ncde/cancersurvival/tools/index.html> (accessed Nov 20, 2014).
- 14 Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002; **21**: 2175–97.
- 15 Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med* 2007; **26**: 5486–98.
- 16 Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974; **19**: 716–23.
- 17 Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata J* 2009; **9**: 265–90.
- 18 StataCorp. STATA statistical software. Texas: Stata Corporation, 2013.
- 19 Department of Health. The NHS Cancer Plan: a plan for investment, a plan for reform. London: Department of Health, 2000.
- 20 Perme MP, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012; **68**: 113–20.
- 21 Office for National Statistics. Cancer statistics: registrations of cancer diagnosed in 2011, England. Series MB1 no. 42. Newport: Office for National Statistics, 2013.
- 22 Møller H, Richards S, Hanchett N, et al. Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates. *Br J Cancer* 2011; **105**: 170–76.
- 23 Swerdlow AJ, Douglas AJ, Vaughan Hudson G, Vaughan Hudson B. Completeness of cancer registration in England and Wales: an assessment based on 2,145 patients with Hodgkin's disease independently registered by the British National Lymphoma Investigation. *Br J Cancer* 1993; **67**: 326–29.

- 24 Coleman MP, Babb P, Damiecki P, et al. Cancer survival trends in England and Wales 1971–1995: deprivation and NHS region. (studies on medical and population subjects no. 61). London: The Stationery Office, 1999.
- 25 Coleman MP, Rachet B, Woods LM, et al. Trends and socioeconomic inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004; **90**: 1367–73.
- 26 Borland S. Half of cancer patients are living for ten years or more: Number classed as having beaten the disease doubles since the 1970s. 2014. <http://www.dailymail.co.uk/health/article-2615416/Half-cancer-patients-living-ten-years-Number-classed-having-beaten-disease-doubles-1970s.html#ixzz37YTäDOIN> (accessed April 28, 2014).
- 27 Rowan S, Rachet B, Alexe DM, Cooper N, Coleman MP. Survival from prostate cancer in England and Wales up to 2001. *Br J Cancer* 2008; **99** (suppl 1): S75–77.
- 28 Klint A, Tryggvadóttir L, Bray F, et al. Trends in the survival of patients diagnosed with cancer in female genital organs in the Nordic countries 1964–2003 followed up to the end of 2006. *Acta Oncol* 2010; **49**: 632–43.
- 29 Puliti D, Duffy SW, Miccinesi G, et al, and the EUROSREEN Working Group. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. *J Med Screen* 2012; **19** (suppl 1): 42–56.
- 30 Lawrence G, Wallis M, Allgood P, et al. Population estimates of survival in women with screen-detected and symptomatic breast cancer taking account of lead time and length bias. *Breast Cancer Res Treat* 2009; **116**: 179–85.
- 31 Woods LM, Rachet B, Cooper N, Coleman MP. Predicted trends in long-term breast cancer survival in England and Wales. *Br J Cancer* 2007; **96**: 1135–38.
- 32 Woods LM, Rachet B, O'Connell DL, et al. Differences in breast cancer incidence in Australia and England by age, extent of disease and deprivation status: women diagnosed 1980–2002. *Aust N Z J Public Health* 2010; **34**: 206–13.
- 33 Coleman MP, Forman D, Bryant H, et al, and the ICBP Module 1 Working Group. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet* 2011; **377**: 127–38.
- 34 National Institute for Health and Care Excellence. Lung cancer—the diagnosis and treatment of lung cancer. 2011. <https://www.nice.org.uk/guidance/cg121> (accessed Nov 20, 2014)
- 35 Sheldon TA, Cullum N, Dawson D, et al. What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. *BMJ* 2004; **329**: 999–1003.
- 36 Pollock A, Macfarlane A. Opting out of care data is not the answer. 2014. <http://www.opendemocracy.net/ournh/allyson-pollock-alison-macfarlane/opting-out-of-care.data-is-not-answer> (accessed April 16, 2014).

4.6.10 Results 2: Index of cancer survival for CCGs [2–5]

The results of the index of cancer survival for CCGs have not been published in a peer-reviewed journal. Instead, these results have been incorporated in annual technical reports published by the Office for National Statistics [2–5]. A summary of the main results is presented below.

Table 4.7 provides a summary of the number of times each of the 8 candidate models was selected for all the cancer-sex combinations. Model 6, the simplest candidate model, was the most selected for 60.7% of the cancer-sex combinations. Model 5 was the second most selected for 15.4% of the combinations, of which about half (119 models) were selected when modelling the breast cancer dataset.

Candidate model	No. of times selected	%
Model 1	5	0.33
Model 2	75	5.08
Model 3	90	6.1
Model 4	57	3.86
Model 5	227	15.4
Model 6	896	60.7
Model 7	38	2.57
Model 8	89	6.03

Table 4.7: CCG index: summary of best fitting models for all the cancer-sex combinations

For the CCG index, an extra model (Model 8) was added to the set of candidate models, compared to the set of 7 models used for the England index. This was because for a few of the combinations of CCG-sex-cancer none of the first 7 models had converged. Adding an extra model ensured that for each combination, there was at least one model that converged and that could be used in the post-estimation procedure. Even so, post-estimation of net survival produced missing estimates (mainly ‘zero’ estimates) for some of the components needed for the construction of the index. This affected the estimation of net survival for colorectum and lung cancer. For colorectum cancer, 5.8% of net survival estimates were missing, corresponding to 1,974 combinations out of a total of 33,760 combinations of

CCG, sex, cancer, age group and year of diagnosis. For lung cancer, 10.6% of net survival estimates were missing, amounting to 3,569 combinations out the total of 33,760. For both cancers, over 80% of these missing estimates occurred in the youngest age group (15-44 years). The post-estimation procedure was re-run, to obtain an estimate for the missing combinations, using merged age groups of the missing age group with an adjacent (non-missing) age group.

Mid-year population estimates (thousands) for 2011, and number of cancer patients included in survival analyses, by calendar year of diagnosis are presented in Tables 4.8 and 4.9. The average CCG mid-year population estimate was around 250,000 inhabitants in 2011. A total of 2,847,166 patients were included in the analysis, ranging between 2,524 and 54,747 patients by CCGs. Tables 4.10 and 4.11 present the one-year net survival index (%) and associated precisions (prec) for CCGs by calendar year of diagnosis. The box-plots in Figure 4.4 summarise the range of estimates for CCGs by year of diagnosis between 1996-2011, showing that the median net survival index has increased steadily, from 58.7% in 1996 to 67% in 2011 and the range of CCG estimates reduced over the years. However, understanding the overall geographical patterns and the spread of individual CCGs from such long tables of results is challenging. The next chapter will explore data visualisation options to best present these results and improve their interpretation.

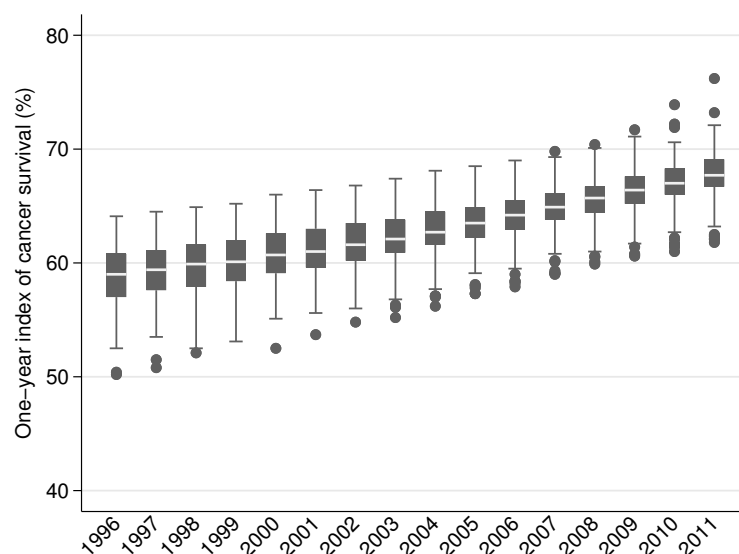


Figure 4.4: Box-plots of one-year net survival index (%) for CCGs by calendar year of diagnosis: all adults, 1996-2011

Table 4.8: Mid-year population estimates (thousands) for 2011, and number of cancer patients included in survival analyses, by calendar year of diagnosis 1996-2003: Clinical Commissioning Groups by regions, England

Clinical Commissioning Group	Year of diagnosis									
	Population	1996	1997	1998	1999	2000	2001	2002	2003	
England	53,107,169	161,929	164,749	165,413	168,248	170,562	171,381	171,573	173,467	
North of England	15,086,775	52,529	52,419	52,993	52,354	54,212	54,384	54,388	53,927	
Eastern Cheshire	194,793	663	669	776	624	742	693	744	640	
South Cheshire	175,943	571	591	587	528	545	544	519	478	
Vale Royal	102,144	296	323	334	331	329	339	301	296	
Warrington	202,709	666	619	649	630	582	549	588	467	
West Cheshire	227,382	839	682	710	837	848	779	848	722	
Wirral	319,837	1,356	1,345	1,298	1,118	1,169	1,255	1,264	1,021	
Darlington	105,584	395	347	439	402	410	431	370	418	
Durham Dales, Easington and Sedgefield	272,878	1,095	1,109	1,060	1,079	1,107	1,076	1,175	1,087	
Hartlepool and Stockton-on-Tees	283,912	964	1,006	963	994	1,032	999	1,061	1,110	
North Durham	240,116	859	833	829	800	863	921	906	932	
South Tees	273,532	1,061	1,115	1,058	1,005	1,093	1,076	1,161	1,105	
Bolton	277,296	882	904	879	890	919	885	777	887	
Bury	185,422	623	647	595	625	617	646	619	625	
Central Manchester	179,709	437	457	431	395	386	371	395	395	
Heywood, Middleton and Rochdale	211,929	731	704	693	733	698	717	684	687	
North Manchester	163,371	563	557	520	507	504	509	435	474	
Oldham	225,157	794	715	719	718	751	751	722	713	
Salford	234,487	969	862	846	826	728	768	596	835	

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
South Manchester	159,822	556	554	546	522	529	509	479	463		
Stockport	283,253	1,078	1,056	1,056	1,038	1,021	1,022	987	1,021		
Tameside and Glossop	252,885	894	894	916	847	853	919	860	860		
Trafford	227,091	864	803	779	789	783	793	737	778		
Wigan Borough	318,122	994	1,003	1,090	1,097	1,018	1,078	919	1,056		
Blackburn with Darwen	147,657	456	447	415	395	457	430	436	433		
Blackpool	142,080	686	637	641	615	599	664	614	612		
Chorley and South Ribble	166,457	505	527	479	556	589	477	540	562		
East Lancashire	371,291	1,250	1,244	1,197	1,237	1,192	1,395	1,258	1,237		
Fylde and Wyre	165,101	717	749	738	746	693	725	753	730		
Greater Preston	201,580	611	601	624	648	682	669	652	660		
Lancashire North	156,512	558	564	569	555	525	590	586	581		
West Lancashire	110,617	396	378	389	392	409	356	376	380		
Halton	125,722	419	424	425	375	391	397	422	361		
Knowsley	145,903	554	515	582	568	545	508	578	470		
Liverpool	465,656	1,861	1,791	1,765	1,721	1,655	1,728	1,758	1,481		
South Sefton	159,764	725	640	641	652	622	668	645	529		
Southport and Formby	114,205	539	532	512	493	444	459	494	369		
St Helens	175,405	657	638	708	543	536	623	702	583		
Cumbria	505,902	1,922	1,889	1,889	1,831	2,026	1,960	2,048	2,089		
Gateshead	200,349	804	801	801	819	875	868	868	866		
Newcastle North and East	139,067	435	441	502	491	531	546	511	538		
Newcastle West	140,025	529	523	548	552	580	612	578	568		

Clinical Commissioning Group	Population	Year of diagnosis										
		1996	1997	1998	1999	2000	2001	2002	2003			
North Tyneside	201,206	793	750	841	824	833	922	862	910			
Northumberland	316,278	1,041	1,024	1,182	1,272	1,287	1,290	1,344	1,336			
South Tyneside	148,164	688	638	689	677	732	669	686	688			
Sunderland	275,330	996	1,011	1,175	1,105	1,153	1,144	1,191	1,121			
East Riding of Yorkshire	313,386	1,075	1,074	1,159	1,078	1,281	1,253	1,253	1,323			
Hambleton, Richmondshire and Whitby	152,737	450	521	477	476	564	541	581	541			
Harrogate and Rural District	158,683	511	498	567	561	573	584	596	629			
Hull	256,123	914	867	886	956	908	922	1,010	977			
North East Lincolnshire	159,735	570	537	557	548	599	591	610	563			
North Lincolnshire	167,516	568	546	548	558	604	592	617	636			
Scarborough and Ryedale	110,351	423	467	426	442	512	458	498	494			
Vale of York	343,046	1,064	1,067	1,105	1,089	1,225	1,170	1,219	1,250			
Barnsley	231,865	825	817	851	820	844	811	850	867			
Bassetlaw	113,003	337	364	379	425	375	361	412	430			
Doncaster	302,468	1,063	1,039	1,051	1,094	1,125	1,083	1,130	1,145			
Rotherham	257,716	828	882	803	861	919	901	925	919			
Sheffield	551,756	1,743	1,977	1,718	1,772	1,924	1,841	1,708	1,885			
Airedale, Wharfedale and Craven	158,328	572	523	585	561	604	626	558	576			
Bradford City	81,741	154	170	167	149	163	176	164	162			
Bradford Districts	332,420	929	976	993	1,005	1,071	1,023	1,067	1,143			
Calderdale	204,170	633	622	597	600	665	729	732	674			
Greater Huddersfield	237,536	661	727	703	668	774	774	784	767			
Leeds North	198,724	647	774	782	741	842	757	810	821			

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
Leeds South and East	235,540	755	791	817	876	902	903	869	906		
Leeds West	316,419	958	993	1,034	977	1,023	1,022	1,085	1,107		
North Kirklees	185,434	529	556	532	568	609	638	619	611		
Wakefield	326,433	1,028	1,072	1,171	1,127	1,223	1,298	1,242	1,327		
Midlands and East of England	16,117,771	48,402	48,603	49,726	50,740	50,765	50,765	51,267	52,466		
Coventry and Rugby	417,411	1,269	1,332	1,279	1,275	1,239	1,225	1,220	1,185		
Herefordshire	183,619	567	651	573	609	610	578	589	643		
Redditch and Bromsgrove	178,050	559	545	557	493	547	512	530	528		
South Warwickshire	258,560	799	860	812	816	769	846	798	838		
South Worcestershire	290,459	932	934	921	897	890	870	842	954		
Warwickshire North	187,498	581	605	585	587	585	610	605	588		
Wyre Forest	98,048	339	345	310	339	356	311	270	329		
Birmingham CrossCity	714,410	2,286	2,219	2,348	2,237	2,158	2,130	2,003	1,947		
Birmingham South and Central	198,331	570	542	572	577	489	470	477	508		
Dudley	313,261	1,038	1,101	1,134	1,072	1,153	1,079	1,082	1,074		
Sandwell and West Birmingham	470,584	1,358	1,335	1,405	1,297	1,308	1,262	1,218	1,207		
Solihull	206,856	661	719	757	778	736	727	678	684		
Walsall	269,524	832	886	888	870	896	894	871	820		
Wolverhampton	249,852	771	820	864	853	798	774	841	812		
Erewash	94,230	331	301	289	329	302	280	294	327		
Hardwick	108,262	369	419	347	413	366	380	428	456		
Mansfield and Ashfield	191,956	630	634	595	594	644	653	572	658		
Newark and Sherwood	114,985	366	354	367	318	366	402	339	422		

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
North Derbyshire	271,899	974	917	928	991	985	994	1,045	1,090		
Nottingham City	303,899	942	962	848	839	860	852	811	779		
Nottingham North and East	145,855	482	474	497	465	457	476	475	473		
Nottingham West	109,749	385	370	374	379	399	315	358	376		
Rushcliffe	111,248	322	315	340	369	311	337	308	320		
Southern Derbyshire	512,082	1,512	1,612	1,595	1,717	1,603	1,686	1,624	1,732		
Cambridgeshire and Peterborough	840,855	2,223	2,234	2,324	2,287	2,459	2,495	2,591	2,717		
Great Yarmouth and Waveney	212,780	737	772	792	857	816	890	870	1,038		
Ipswich and East Suffolk	394,882	1,216	1,246	1,312	1,346	1,364	1,327	1,437	1,545		
North Norfolk	167,524	598	639	654	717	715	701	778	806		
Norwich	191,038	650	613	583	638	612	650	722	714		
South Norfolk	232,895	717	718	778	786	842	783	921	900		
West Norfolk	170,545	574	640	635	650	679	693	734	731		
West Suffolk	219,895	642	645	655	713	671	684	805	815		
Basildon and Brentwood	248,812	718	706	781	740	769	795	858	869		
Castle Point, Rayleigh and Rochford	171,297	545	555	571	620	615	578	629	636		
Mid Essex	377,725	909	949	949	1,006	1,175	1,084	1,077	1,141		
North East Essex	311,676	1,024	954	1,006	1,148	1,151	1,148	1,208	1,173		
Southend	174,274	638	562	654	640	624	569	675	605		
Thurrock	158,268	383	387	397	384	410	485	485	434		
West Essex	287,089	719	673	839	896	831	899	854	924		
Bedfordshire	413,484	1,108	1,193	1,325	1,280	1,283	1,332	1,336	1,377		
Corby	61,607	175	182	193	203	188	195	211	191		

Clinical Commissioning Group	Population	Year of diagnosis										
		1996	1997	1998	1999	2000	2001	2002	2003			
East and North Hertfordshire	535,855	1,361	1,449	1,544	1,550	1,632	1,540	1,420	1,622			
Herts Valleys	565,499	1,552	1,565	1,500	1,573	1,627	1,590	1,594	1,491			
Luton	203,641	489	499	506	522	470	592	531	519			
Milton Keynes	255,399	536	531	538	628	593	639	608	616			
Nene	616,744	1,887	1,691	1,852	1,896	1,909	1,817	1,894	1,966			
East Leicestershire and Rutland	318,516	847	873	843	948	972	942	1,041	1,041			
Leicester City	329,627	769	754	751	786	829	864	816	822			
Lincolnshire East	227,771	830	810	925	872	769	888	883	934			
Lincolnshire West	225,253	704	709	689	702	741	688	793	763			
South Lincolnshire	140,465	431	429	436	455	431	472	507	479			
South West Lincolnshire	121,279	340	332	305	361	382	388	395	413			
West Leicestershire	370,244	1,010	1,034	1,063	1,105	1,180	1,103	1,167	1,212			
Cannock Chase	132,287	405	389	430	435	468	386	457	443			
East Staffordshire	123,312	365	366	395	336	382	365	356	346			
North Staffordshire	212,906	772	742	741	789	758	779	750	783			
Shropshire	307,108	1,121	991	1,019	1,107	1,055	1,096	1,017	1,051			
South East Staffs and Seisdon and Peninsular	222,365	697	651	673	695	651	736	682	675			
Stafford and Surrounds	150,495	458	529	523	489	486	533	523	529			
Stoke on Trent	256,900	928	932	923	986	921	931	919	931			
Telford and Wrekin	166,831	449	407	437	480	478	445	445	464			
London	8,204,407	18,611	19,214	19,533	19,676	20,417	19,477	19,596	19,971			
Barking and Dagenham	187,029	553	564	540	579	586	532	558	526			
Barnet	357,538	854	816	885	888	933	863	868	966			

Clinical Commissioning Group	Population	Year of diagnosis										
		1996	1997	1998	1999	2000	2001	2002	2003			
Bexley	232,774	665	770	658	736	765	749	721	728			
Brent	312,245	594	595	626	607	616	608	646	607			
Bromley	310,554	956	945	998	1,019	1,092	1,034	1,035	1,062			
Camden	220,087	540	532	538	540	538	536	507	550			
Central London (Westminster)	157,640	370	356	392	363	398	348	405	394			
City and Hackney	254,594	483	447	435	480	501	513	491	491			
Croydon	364,815	863	940	900	961	996	908	1,003	936			
Ealing	339,314	642	693	734	773	710	710	656	794			
Enfield	313,935	787	744	769	750	768	768	737	767			
Greenwich	255,483	613	656	621	626	664	635	700	697			
Hammersmith and Fulham	182,445	312	352	393	378	419	391	338	429			
Haringey	255,540	472	471	492	488	501	446	522	530			
Harrow	240,499	578	576	557	567	637	534	571	598			
Havering	237,927	693	764	893	836	872	833	784	882			
Hillingdon	275,499	695	694	692	687	673	641	646	584			
Hounslow	254,927	425	510	541	504	476	492	499	546			
Islington	206,285	455	498	483	492	538	502	509	495			
Kingston	160,436	397	424	446	403	502	479	479	437			
Lambeth	304,481	552	593	586	576	645	613	668	631			
Lewisham	276,938	605	677	644	689	654	654	642	604			
Merton	200,543	535	584	618	539	594	530	490	599			
Newham	310,460	495	513	546	492	504	544	504	485			
Redbridge	281,395	682	637	682	621	666	689	639	700			

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
Richmond	187,527	462	536	517	527	506	476	562	559		
Southwark	288,717	531	609	591	693	690	629	666	609		
Sutton	191,123	548	598	590	587	626	643	573	560		
Tower Hamlets	256,012	476	362	422	465	465	462	471	456		
Waltham Forest	259,742	589	554	542	597	605	531	508	545		
Wandsworth	307,710	775	719	740	710	735	685	710	676		
West London	220,193	414	485	462	503	542	499	488	528		
South of England	13,698,216	42,387	44,513	43,161	45,478	45,168	46,755	46,322	47,103		
Bath and North East Somerset	175,538	547	791	636	610	634	656	619	649		
Gloucestershire	598,289	1,754	1,954	1,900	2,084	1,961	2,026	1,862	2,100		
Swindon	214,944	579	557	490	620	530	600	516	590		
Wiltshire	474,319	1,390	1,436	1,266	1,410	1,470	1,537	1,519	1,551		
Bristol	428,074	1,159	1,423	1,364	1,342	1,191	1,335	1,273	1,410		
North Somerset	203,091	583	747	707	708	719	780	804	783		
Somerset	531,581	1,672	1,860	1,768	2,191	1,940	2,086	1,950	2,018		
South Gloucestershire	263,417	658	759	742	865	717	761	751	885		
Kernow	535,984	1,857	1,987	1,872	2,026	1,962	2,156	2,155	2,182		
North East West Devon	863,433	2,797	3,283	3,317	3,056	3,081	3,478	3,270	3,459		
South Devon and Torbay	272,058	1,011	1,219	1,024	1,268	1,172	1,242	1,213	1,158		
Ashford	118,405	287	326	254	340	301	327	321	322		
Canterbury and Coastal	197,807	557	548	585	639	642	609	617	647		
Dartford, Gravesham and Swanley	246,390	649	621	695	675	756	802	728	787		
Medway	264,885	609	799	747	665	776	762	703	791		

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
South Kent Coast	202,193	551	620	650	737	681	702	813	741		
Swale	106,841	279	255	271	332	303	346	355	315		
Thanet	134,402	405	444	572	533	624	574	564	560		
West Kent	460,428	1,177	1,197	1,269	1,334	1,366	1,352	1,362	1,390		
Brighton and Hove	272,952	802	697	788	858	813	827	843	753		
Coastal West Sussex	473,254	1,793	1,787	1,922	1,886	1,895	2,062	2,140	1,985		
Crawley	107,053	242	298	280	267	312	264	258	281		
East Surrey	174,374	480	491	500	520	505	589	532	544		
Eastbourne, Hailsham and Seaford	180,397	715	716	765	797	798	825	784	736		
Guildford and Waverley	204,102	619	572	647	613	664	630	634	641		
Hastings and Rother	180,902	576	667	702	705	696	686	695	730		
High Weald Lewes Havens	165,910	438	494	510	528	575	515	545	595		
Horsham and Mid Sussex	221,833	572	535	641	629	676	685	643	619		
North West Surrey	336,391	945	1,039	973	924	982	956	966	975		
Surrey Downs	280,770	909	980	992	962	1,029	929	1,012	1,053		
Surrey Heath	93,513	243	286	266	308	246	277	275	277		
Aylesbury Vale	193,274	564	546	567	597	599	508	614	583		
Bracknell and Ascot	131,791	367	361	366	397	361	331	305	320		
Chiltern	316,094	1,034	997	1,035	1,016	1,006	1,042	967	1,028		
Newbury and District	104,639	300	301	308	288	289	306	283	323		
North and West Reading	99,330	277	310	279	321	288	319	325	356		
Oxfordshire	641,234	1,877	1,840	1,955	2,025	1,915	1,957	2,044	1,971		
Slough	140,713	311	316	311	311	297	303	310	345		

Clinical Commissioning Group	Population	Year of diagnosis									
		1996	1997	1998	1999	2000	2001	2002	2003		
South Reading	105,518	246	237	233	291	236	241	252	269		
Windsor, Ascot and Maidenhead	137,584	436	399	444	457	442	453	406	464		
Wokingham	154,943	428	395	426	437	442	441	453	480		
Dorset	745,338	3,130	3,069	2,709	2,955	2,951	3,131	3,117	3,213		
Fareham and Gosport	194,600	704	578	420	558	604	663	699	666		
Isle of Wight	138,392	623	582	454	503	541	530	513	631		
North East Hampshire and Farnham	205,729	415	701	520	508	594	572	594	566		
North Hampshire	214,038	626	538	563	478	569	565	627	551		
Portsmouth	205,433	756	733	516	710	572	657	677	598		
South Eastern Hampshire	208,475	786	701	663	765	745	720	783	736		
Southampton	235,870	727	705	691	659	762	716	734	630		
West Hampshire	541,691	1,925	1,816	1,586	1,770	1,938	1,924	1,897	1,846		

Table 4.9: Mid-year population estimates (thousands) for 2011, and number of cancer patients included in survival analyses, by calendar year of diagnosis 2004-2011: Clinical Commissioning Groups by regions, England

England	175,716	178,203	183,504	184,243	191,661	195,099	196,816	194,602	2,847,166
North of England	54,036	54,341	55,509	53,816	56,718	59,897	59,760	60,131	881,414
Eastern Cheshire	652	748	768	539	636	880	872	827	11,473
South Cheshire	583	507	523	388	573	735	774	703	9,149
Vale Royal	285	284	368	229	331	434	419	409	5,308
Warrington	561	566	625	533	647	796	805	734	10,017
West Cheshire	742	834	790	835	951	1,002	1,075	965	13,459
Wirral	1,073	1,238	1,135	1,176	1,396	1,428	1,444	1,388	20,104
Darlington	362	392	371	370	450	402	397	428	6,384
Durham Dales, Easington and Sedgfield	1,074	1,117	1,111	1,087	1,206	1,223	1,147	1,154	17,907
Hartlepool and Stockton-on-Tees	1,101	1,051	1,067	1,101	1,138	1,204	1,125	1,158	17,074
North Durham	931	908	927	887	965	933	871	967	14,332
South Tees	1,137	1,038	1,071	1,144	1,104	1,166	1,177	1,214	17,725
Bolton	857	891	964	866	853	1,054	1,033	959	14,500
Bury	636	641	699	615	611	695	784	701	10,379
Central Manchester	380	392	407	343	416	429	442	366	6,442
Heywood, Middleton and Rochdale	685	703	730	602	681	761	804	759	11,372
North Manchester	494	479	498	444	445	547	495	447	7,918
Oldham	708	668	784	691	687	821	849	772	11,863
Salford	802	824	910	851	870	886	890	907	13,370
South Manchester	484	498	547	487	472	487	540	541	8,214
Stockport	1,005	1,064	1,242	1,075	1,026	1,146	1,165	1,196	17,198
Tameside and Glossop	854	859	914	916	878	948	1,020	961	14,393

Clinical Commissioning Group	Year of diagnosis										Total
	2004	2005	2006	2007	2008	2009	2010	2011			
Trafford	790	786	919	771	826	861	872	804	12,955		
Wigan Borough	1,059	1,074	1,200	1,025	1,056	1,252	1,273	1,212	17,406		
Blackburn with Darwen	400	451	473	382	443	498	452	466	7,034		
Blackpool	640	691	649	597	583	655	762	646	10,291		
Chorley and South Ribble	537	588	581	612	535	655	632	652	9,027		
East Lancashire	1,300	1,268	1,257	1,253	1,369	1,560	1,359	1,363	20,739		
Fylde and Wyre	766	825	812	755	731	912	827	846	12,325		
Greater Preston	686	652	687	700	629	755	783	739	10,778		
Lancashire North	566	548	577	599	674	701	686	693	9,572		
West Lancashire	317	351	356	280	470	478	458	476	6,262		
Halton	343	420	419	418	495	550	519	493	6,871		
Knowsley	522	548	525	574	643	668	698	629	9,127		
Liverpool	1,550	1,606	1,617	1,622	1,776	1,911	1,987	1,934	27,763		
South Sefton	586	603	589	616	808	777	782	771	10,654		
Southport and Formby	388	443	484	410	556	570	575	595	7,863		
St Helens	557	606	697	595	645	765	731	782	10,368		
Cumbria	2,090	2,069	2,128	2,158	2,135	2,177	2,116	2,226	32,753		
Gateshead	864	838	803	790	890	822	855	877	13,441		
Newcastle North and East	524	495	480	504	430	471	472	489	7,860		
Newcastle West	557	524	575	591	550	506	557	512	8,862		
North Tyneside	847	823	833	860	792	849	857	852	13,448		
Northumberland	1,382	1,322	1,307	1,335	1,333	1,388	1,333	1,476	20,652		
South Tyneside	735	687	722	677	718	660	682	665	11,013		

Clinical Commissioning Group	Year of diagnosis										Total
	2004	2005	2006	2007	2008	2009	2010	2011			
Sunderland	1,135	1,073	1,105	1,171	1,155	1,265	1,182	1,174	18,156		
East Riding of Yorkshire	1,302	1,295	1,294	1,338	1,336	1,376	1,436	1,412	20,285		
Hambleton, Richmondshire and Whitby	554	588	541	543	609	542	608	647	8,783		
Harrogate and Rural District	629	557	548	621	561	570	683	591	9,279		
Hull	1,002	899	951	957	969	999	884	948	15,049		
North East Lincolnshire	594	584	559	607	646	648	619	662	9,494		
North Lincolnshire	599	620	609	663	684	686	675	676	9,881		
Scarborough and Ryedale	487	451	523	444	491	565	469	497	7,647		
Vale of York	1,219	1,153	1,258	1,156	1,253	1,238	1,229	1,353	19,048		
Barnsley	883	887	932	951	967	881	938	929	14,053		
Bassetlaw	435	452	406	422	464	493	449	500	6,704		
Doncaster	1,175	1,206	1,162	1,163	1,206	1,322	1,225	1,270	18,459		
Rotherham	1,025	990	1,013	998	1,042	1,020	1,054	1,120	15,300		
Sheffield	1,841	1,972	1,992	1,987	2,074	2,048	1,944	2,081	30,507		
Airedale, Wharfedale and Craven	605	609	592	556	634	521	632	726	9,480		
Bradford City	149	153	135	143	157	146	158	178	2,524		
Bradford Districts	1,060	1,107	1,100	993	1,080	1,108	1,015	1,097	16,767		
Calderdale	711	670	654	745	731	769	711	745	10,988		
Greater Huddersfield	716	768	706	756	813	763	784	789	11,953		
Leeds North	785	732	750	788	710	781	804	803	12,327		
Leeds South and East	860	834	825	791	842	842	880	943	13,636		
Leeds West	1,037	1,114	1,006	982	1,015	1,042	1,042	1,214	16,651		
North Kirklees	572	503	571	526	588	600	625	639	9,286		

Clinical Commissioning Group	Year of diagnosis										Total
	2004	2005	2006	2007	2008	2009	2010	2011			
Wakefield	1,249	1,204	1,136	1,212	1,268	1,254	1,318	1,383	19,512		
Midlands and East of England	53,875	55,442	56,653	58,404	60,951	61,364	62,081	59,762	871,266		
Coventry and Rugby	1,218	1,254	1,200	1,240	1,345	1,353	1,392	1,370	20,396		
Herefordshire	663	643	691	682	718	712	755	705	10,389		
Redditch and Bromsgrove	558	533	538	584	607	623	641	634	8,989		
South Warwickshire	850	779	788	826	879	988	940	928	13,516		
South Worcestershire	1,044	987	946	1,081	1,117	1,104	1,177	1,148	15,844		
Warwickshire North	622	556	584	627	639	699	705	677	9,855		
Wyre Forest	370	325	323	396	385	427	400	413	5,638		
Birmingham CrossCity	2,101	2,078	2,101	2,133	2,183	2,249	2,213	2,233	34,619		
Birmingham South and Central	479	510	523	573	600	550	538	589	8,567		
Dudley	1,157	1,014	996	1,109	1,106	1,167	1,161	1,179	17,622		
Sandwell and West Birmingham	1,192	1,207	1,173	1,304	1,312	1,458	1,394	1,407	20,837		
Solihull	727	733	708	756	764	814	893	858	11,993		
Walsall	974	866	865	869	981	934	944	986	14,376		
Wolverhampton	757	658	797	805	836	904	913	891	13,094		
Erewash	347	368	362	373	321	394	380	332	5,330		
Hardwick	431	435	400	417	441	468	447	439	6,656		
Mansfield and Ashfield	712	688	752	698	755	739	801	771	10,896		
Newark and Sherwood	449	425	479	464	471	472	533	479	6,706		
North Derbyshire	1,068	1,202	1,141	1,067	1,092	1,096	1,158	1,217	16,965		
Nottingham City	856	883	917	904	977	898	886	919	14,133		
Nottingham North and East	533	585	550	586	612	582	593	621	8,461		

Clinical Commissioning Group	Year of diagnosis									Total
	2004	2005	2006	2007	2008	2009	2010	2011		
Nottingham West	435	424	426	460	445	437	440	442	6,465	
Rushcliffe	395	413	367	417	425	443	437	450	5,969	
Southern Derbyshire	1,701	1,689	1,846	1,847	1,825	1,808	1,865	1,933	27,595	
Cambridgeshire and Peterborough	2,789	2,916	2,961	3,043	3,254	3,268	3,278	3,028	43,867	
Great Yarmouth and Waveney	972	941	1,118	1,115	1,154	1,120	1,163	1,091	15,446	
Ipswich and East Suffolk	1,558	1,556	1,582	1,610	1,668	1,754	1,741	1,625	23,887	
North Norfolk	774	841	909	861	951	902	974	834	12,654	
Norwich	772	765	782	711	830	795	843	690	11,370	
South Norfolk	952	987	1,053	1,050	1,122	1,139	1,151	1,050	14,949	
West Norfolk	732	770	841	826	966	970	914	841	12,196	
West Suffolk	822	864	876	888	1,009	994	994	905	12,982	
Basildon and Brentwood	859	879	902	883	952	1,051	978	831	13,571	
Castle Point, Rayleigh and Rochford	631	727	767	778	823	822	823	819	10,939	
Mid Essex	1,106	1,369	1,417	1,419	1,465	1,536	1,585	1,522	19,709	
North East Essex	1,336	1,462	1,408	1,466	1,514	1,485	1,524	1,460	20,467	
Southend	685	678	764	744	736	736	813	673	10,796	
Thurrock	408	475	494	551	519	560	623	495	7,490	
West Essex	888	1,144	1,092	1,158	1,285	1,236	1,166	992	15,596	
Bedfordshire	1,445	1,336	1,532	1,538	1,594	1,628	1,600	1,350	22,257	
Corby	225	241	217	217	250	213	238	251	3,390	
East and North Hertfordshire	1,484	1,838	1,890	1,968	2,182	2,124	1,981	1,805	27,390	
Herts Valleys	1,403	1,956	1,919	1,946	2,013	2,009	2,032	1,756	27,526	
Luton	583	564	593	603	582	593	636	569	8,851	

Clinical Commissioning Group	Year of diagnosis									Total
	2004	2005	2006	2007	2008	2009	2010	2011		
Milton Keynes	729	684	705	730	788	739	802	750	10,616	
Nene	2,050	2,001	2,127	2,169	2,217	2,091	2,055	2,121	31,743	
East Leicestershire and Rutland	1,076	1,176	1,055	1,133	1,302	1,125	1,165	1,344	16,883	
Leicester City	790	818	893	896	811	888	842	844	13,173	
Lincolnshire East	1,028	952	1,018	1,099	1,163	1,086	1,099	1,167	15,523	
Lincolnshire West	713	848	866	837	957	862	917	922	12,711	
South Lincolnshire	508	504	515	595	629	596	614	600	8,201	
South West Lincolnshire	430	412	506	537	508	525	492	486	6,812	
West Leicestershire	1,196	1,225	1,270	1,300	1,280	1,349	1,434	1,348	19,276	
Cannock Chase	390	417	419	415	454	493	474	499	6,974	
East Staffordshire	343	388	367	352	402	400	448	416	6,027	
North Staffordshire	748	745	716	759	774	834	806	831	12,327	
Shropshire	1,122	989	1,055	1,146	1,094	1,191	1,272	1,252	17,578	
South East Staffs and Seisdon and Peninsular	713	761	695	830	754	785	868	806	11,672	
Stafford and Surrounds	502	541	516	471	587	589	573	589	8,438	
Stoke on Trent	1,005	900	851	1,006	958	994	1,009	976	15,170	
Telford and Wrekin	469	517	489	536	568	563	548	603	7,898	
London	19,475	20,276	20,665	21,258	21,586	21,998	21,759	21,734	325,246	
Barking and Dagenham	483	512	550	505	554	508	482	507	8,539	
Barnet	924	922	962	1,027	1,078	1,099	1,038	1,096	15,219	
Bexley	757	788	777	812	784	866	838	793	12,207	
Brent	577	663	630	632	668	712	737	773	10,291	
Bromley	1,006	1,050	1,166	1,084	1,139	1,171	1,127	1,142	17,026	

Clinical Commissioning Group	Year of diagnosis									Total
	2004	2005	2006	2007	2008	2009	2010	2011		
Camden	541	499	517	586	552	588	599	584	8,747	
Central London (Westminster)	340	372	375	373	412	354	379	360	5,991	
City and Hackney	456	530	527	482	521	575	517	512	7,961	
Croydon	953	947	1,005	1,045	1,051	1,021	1,034	1,105	15,668	
Ealing	701	811	790	837	768	844	806	775	12,044	
Enfield	756	828	771	807	865	912	871	906	12,806	
Greenwich	649	691	608	605	665	697	675	642	10,444	
Hammersmith and Fulham	386	399	427	437	424	452	476	476	6,489	
Haringey	534	472	543	561	574	563	622	606	8,397	
Harrow	581	615	578	577	614	661	619	685	9,548	
Havering	838	821	825	953	943	896	962	838	13,633	
Hillingdon	687	690	740	807	763	814	803	795	11,411	
Hounslow	528	574	603	641	592	638	688	660	8,917	
Islington	440	483	486	500	535	532	544	545	8,037	
Kingston	446	482	450	511	523	462	464	452	7,357	
Lambeth	643	676	683	677	699	717	659	702	10,320	
Lewisham	671	683	665	678	754	737	709	707	10,773	
Merton	558	528	517	579	560	535	533	583	8,882	
Newham	515	540	528	582	550	581	544	589	8,512	
Redbridge	639	640	701	739	726	732	738	745	10,976	
Richmond	545	557	597	578	579	619	552	592	8,764	
Southwark	663	652	670	688	623	662	679	609	10,264	
Sutton	567	531	593	586	628	700	658	635	9,623	

Clinical Commissioning Group	Year of diagnosis										Total
	2004	2005	2006	2007	2008	2009	2010	2011			
Tower Hamlets	428	468	497	468	468	472	481	409	7,270		
Waltham Forest	531	555	614	569	680	636	653	613	9,322		
Wandsworth	698	721	756	764	719	704	736	768	11,616		
West London	434	576	514	568	575	538	536	530	8,192		
South of England	48,330	48,144	50,677	50,765	52,406	51,840	53,216	52,975	769,240		
Bath and North East Somerset	714	620	591	723	650	632	709	578	10,359		
Gloucestershire	2,120	2,041	2,253	2,276	2,375	2,249	2,393	2,381	33,729		
Swindon	640	640	669	633	621	683	719	653	9,740		
Wiltshire	1,631	1,635	1,668	1,630	1,817	1,680	1,838	1,860	25,338		
Bristol	1,410	1,246	1,388	1,330	1,424	1,457	1,374	1,462	21,588		
North Somerset	848	787	740	839	831	847	864	867	12,454		
Somerset	2,129	2,084	2,204	2,198	2,323	2,324	2,277	2,309	33,333		
South Gloucestershire	912	863	921	952	876	975	1,047	964	13,648		
Kernow	2,229	2,169	2,392	2,447	2,412	2,328	2,466	2,390	35,030		
North East West Devon	3,501	3,442	3,635	3,579	3,678	3,611	3,690	3,870	54,747		
South Devon and Torbay	1,240	1,169	1,294	1,276	1,402	1,309	1,322	1,325	19,644		
Ashford	349	357	432	371	382	419	399	430	5,617		
Canterbury and Coastal	684	742	802	699	740	782	738	782	10,813		
Dartford, Gravesham and Swanley	759	757	850	901	964	835	845	842	12,466		
Medway	858	819	853	849	810	916	899	910	12,766		
South Kent Coast	840	791	827	850	827	857	921	889	12,297		
Swale	299	335	330	336	438	363	410	380	5,347		
Thanet	590	601	534	597	610	551	681	693	9,133		

Clinical Commissioning Group	Year of diagnosis								Total
	2004	2005	2006	2007	2008	2009	2010	2011	
West Kent	1,516	1,426	1,494	1,579	1,655	1,589	1,613	1,557	22,876
Brighton and Hove	803	861	793	836	837	886	896	888	13,181
Coastal West Sussex	2,016	2,029	2,050	2,036	2,210	2,042	2,062	2,178	32,093
Crawley	292	289	337	294	316	322	316	305	4,673
East Surrey	557	496	568	551	579	566	610	577	8,665
Eastbourne, Hailsham and Seaford	811	816	789	799	826	830	890	860	12,757
Guildford and Waverley	594	591	665	633	629	695	729	628	10,184
Hastings and Rother	663	734	819	834	885	848	898	815	11,953
High Weald Lewes Havens	549	534	630	547	590	683	750	674	9,157
Horsham and Mid Sussex	658	674	647	768	712	690	771	795	10,715
North West Surrey	1,061	1,094	1,150	1,125	1,139	1,182	1,210	1,177	16,898
Surrey Downs	945	1,046	965	971	1,014	1,122	1,129	1,083	16,141
Surrey Heath	288	307	312	303	296	300	340	322	4,646
Aylesbury Vale	653	695	616	647	660	648	661	669	9,827
Bracknell and Ascot	350	387	378	412	403	424	377	370	5,909
Chiltern	1,042	1,061	1,114	1,079	1,137	1,082	1,113	1,031	16,784
Newbury and District	369	360	394	424	354	381	405	312	5,397
North and West Reading	330	308	348	382	303	348	348	342	5,184
Oxfordshire	2,117	2,208	2,312	2,270	2,486	2,283	2,335	2,358	33,953
Slough	322	274	315	337	341	283	339	287	5,002
South Reading	265	257	269	243	255	241	262	260	4,057
Windsor, Ascot and Maidenhead	465	472	513	495	480	474	417	495	7,312
Wokingham	441	498	510	471	551	531	484	530	7,518

Clinical Commissioning Group	Year of diagnosis								
	2004	2005	2006	2007	2008	2009	2010	2011	Total
Dorset	3,054	3,097	3,297	3,213	3,257	3,250	3,339	3,307	50,089
Fareham and Gosport	631	702	768	736	783	820	757	852	10,941
Isle of Wight	589	566	592	625	643	627	625	679	9,323
North East Hampshire and Farnham	586	601	655	702	699	688	730	692	9,823
North Hampshire	575	638	622	651	697	753	745	781	9,979
Portsmouth	646	629	693	725	675	665	695	682	10,629
South Eastern Hampshire	779	763	860	819	836	898	863	921	12,638
Southampton	715	689	730	736	759	707	710	761	11,431
West Hampshire	1,895	1,944	2,089	2,036	2,219	2,164	2,205	2,202	31,456

Table 4.10: One-year net survival index (NS: %) and precision of estimates (prec) for all cancers combined, by calendar year of diagnosis: all adults, Clinical Commissioning Groups, England, 1996-2003

Clinical Commissioning Group	Year of diagnosis											
	1996		1997		1998		1999		2000		2001	
	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec
England	59.2	465.7	59.7	653.2	60.0	895.4	60.4	1151.3	60.8	1326.8	61.2	1370.0
North of England												
Eastern Cheshire	61.4	2.6	61.6	3.4	62.2	4.6	62.8	5.5	62.6	6.2	63.0	6.7
South Cheshire	60.2	2.5	59.4	3.3	58.3	4.2	58.4	5.7	57.3	6.1	56.8	5.9
Vale Royal	56.4	2.3	57.1	2.6	57.5	3.1	57.2	3.4	58.1	4.1	58.5	4.6
Warrington	58.8	<0.001	59.3	<0.001	59.6	<0.001	59.9	<0.001	60.3	<0.001	60.7	<0.001
West Cheshire	58.7	<0.001	58.8	<0.001	59.2	<0.001	59.4	<0.001	60.1	<0.001	60.4	<0.001
Wirral	59.1	<0.001	59.3	<0.001	59.6	<0.001	60.0	<0.001	60.4	<0.001	60.7	<0.001
Darlington	60.9	1.9	60.6	2.4	60.8	3.0	61.2	4.0	61.3	4.2	61.9	4.4
Durham Dales, Easington and Sedgefield	57.6	5.6	58.3	7.0	58.6	8.5	59.2	10.2	59.8	11.7	60.5	13.2
Hartlepool and Stockton-on-Tees	57.0	3.5	57.4	4.6	58.3	6.3	58.7	7.8	59.4	9.0	59.9	9.7
North Durham	58.6	<0.001	59.4	<0.001	59.6	<0.001	60.1	<0.001	60.6	<0.001	61.0	<0.001
South Tees	57.8	4.6	58.1	6.4	58.6	8.3	59.4	10.4	59.8	11.2	60.4	11.7
Bolton	55.0	5.1	55.8	6.1	57.0	7.5	57.7	8.8	58.6	10.1	59.5	10.8
Bury	55.2	<0.001	56.0	<0.001	56.6	<0.001	57.7	<0.001	59.0	<0.001	59.9	<0.001
Central Manchester	52.8	1.6	53.6	2.1	54.2	2.7	54.8	3.4	55.4	3.8	56.4	4.0
Heywood, Middleton and Rochdale	54.1	<0.001	55.1	<0.001	56.1	<0.001	57.0	<0.001	58.1	<0.001	58.3	<0.001
North Manchester	50.4	2.2	51.5	3.0	52.5	3.6	54.1	4.5	55.1	4.5	56.3	4.7
Oldham	54.2	4.1	54.8	5.3	55.5	6.3	56.7	7.7	57.2	8.3	57.9	9.2

Clinical Commissioning Group	Year of diagnosis											
	1996		1997		1998		1999		2000		2001	
	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec
Calderdale	56.6	2.3	57.7	3.3	58.5	4.4	59.3	5.5	59.9	6.1	60.3	6.3
Greater Huddersfield	61.1	4.7	61.9	5.6	62.1	6.3	62.5	7.6	62.6	8.5	63.1	9.2
Leeds North	63.7	5.0	64.3	5.9	64.7	7.3	65.1	8.6	65.7	9.9	66.0	10.7
Leeds South and East	60.7	4.9	61.1	5.6	61.6	7.2	61.9	7.9	62.2	9.0	62.7	10.5
Leeds West	61.2	6.5	61.5	7.7	62.3	9.2	62.7	10.2	63.3	12.0	64.0	13.1
North Kirklees	59.0	3.0	59.6	4.0	60.4	4.8	61.3	6.0	61.9	6.9	62.2	7.5
Wakefield	59.2	7.2	60.1	8.6	60.5	9.9	61.1	11.6	61.7	13.7	62.2	15.1
Midlands and East of England												
Coventry and Rugby	57.9	5.0	58.8	7.2	59.3	9.6	59.8	11.8	60.5	12.6	61.0	11.9
Herefordshire	62.4	3.7	62.8	4.6	62.5	5.4	62.8	6.5	63.4	7.7	63.3	8.3
Redditch and Bromsgrove	63.7	2.3	63.3	3.1	63.9	4.3	64.2	5.2	64.0	5.8	64.2	5.8
South Warwickshire	58.6	<0.001	60.0	<0.001	61.1	<0.001	61.8	<0.001	62.9	<0.001	63.8	<0.001
South Worcestershire	61.8	3.6	62.4	4.8	62.8	6.3	63.0	7.8	63.2	8.7	63.6	8.9
Warwickshire North	59.0	2.6	59.3	3.4	60.0	4.7	60.1	5.6	60.7	6.2	61.2	6.3
Wyre Forest	59.9	1.8	60.8	2.3	61.3	2.6	61.3	3.1	62.5	3.9	62.6	4.3
Birmingham CrossCity	59.5	7.9	60.1	11.3	60.4	15.6	61.0	19.3	61.3	20.6	61.7	19.9
Birmingham South and Central	60.8	2.7	61.0	3.5	60.8	4.4	60.6	5.3	60.4	5.6	60.7	5.4
Dudley	58.5	<0.001	58.8	<0.001	59.3	<0.001	59.8	<0.001	60.2	<0.001	60.5	<0.001
Sandwell and West Birmingham	57.4	5.2	58.0	7.1	57.9	9.0	58.4	10.8	58.9	12.3	59.1	12.4
Solihull	62.6	4.4	63.2	5.2	63.5	6.0	63.9	7.0	64.7	8.4	65.0	9.3
Walsall	57.9	3.0	58.0	4.0	58.6	5.2	59.0	6.6	59.3	7.4	60.0	7.7
Wolverhampton	56.8	3.1	57.2	4.5	57.5	5.9	58.3	7.3	59.1	8.2	59.8	7.8

Table 4.11: One-year net survival index (NS: %) and precision of estimates (prec) for all cancers combined, by calendar year of diagnosis: all adults, Clinical Commissioning Groups, England, 2004-2011

Clinical Commissioning Group	Year of diagnosis											
	2004		2005		2006		2007		2008		2009	
	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec
England	62.9	1331.9	63.6	1423.9	64.3	1540.8	65.0	1586.4	65.8	1441.4	66.6	1155.0
North of England												
Eastern Cheshire	64.0	6.5	64.6	6.8	65.1	7.4	65.4	7.5	65.9	6.8	66.5	5.9
South Cheshire	57.0	5.2	57.3	5.3	58.2	5.6	59.1	5.9	60.6	6.1	62.0	5.2
Vale Royal	59.9	5.7	60.4	5.5	59.8	4.6	60.8	4.5	61.4	4.0	60.7	3.1
Warrington	61.8	<0.001	62.2	<0.001	62.8	<0.001	63.8	<0.001	63.4	<0.001	64.2	<0.001
West Cheshire	62.5	<0.001	63.1	<0.001	64.0	<0.001	65.2	<0.001	66.3	<0.001	67.7	<0.001
Wirral	61.9	<0.001	62.3	<0.001	62.7	<0.001	63.4	<0.001	63.6	<0.001	64.2	<0.001
Darlington	62.1	4.0	62.8	4.3	62.9	4.5	63.4	4.4	63.6	3.8	64.1	3.2
Durham Dales, Easington and Sedgefield	62.0	13.9	62.2	12.6	62.5	12.5	63.3	12.3	63.9	10.4	64.2	8.9
Hartlepool and Stockton-on-Tees	62.3	10.0	62.8	10.3	63.8	11.0	63.9	10.6	65.1	10.1	65.7	8.1
North Durham	62.8	<0.001	63.3	<0.001	64.1	<0.001	64.5	<0.001	65.8	<0.001	66.1	<0.001
South Tees	62.0	10.4	63.2	11.4	63.8	12.2	64.3	12.0	65.3	10.8	65.9	9.1
Bolton	62.0	11.8	63.1	11.9	64.0	11.7	64.6	9.9	65.5	9.1	66.2	8.1
Bury	62.8	<0.001	63.6	<0.001	64.5	<0.001	65.5	<0.001	66.5	<0.001	66.8	<0.001
Central Manchester	59.4	3.9	60.5	3.9	62.2	3.9	64.5	4.1	65.8	3.5	66.4	2.4
Heywood, Middleton and Rochdale	61.6	<0.001	62.7	<0.001	63.4	<0.001	64.2	<0.001	65.7	<0.001	66.6	<0.001
North Manchester	59.8	4.5	60.6	4.6	61.6	4.7	62.7	5.0	63.7	4.3	64.5	3.5
Oldham	60.4	9.0	61.1	8.8	62.1	8.8	63.0	8.8	63.9	7.8	64.5	6.6

Clinical Commissioning Group	2004		2005		2006		2007		2008		2009		2010		2011	
	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec	NS	Prec
Surrey Downs	67.9	17.8	68.3	16.9	68.5	15.5	69.2	14.5	69.6	12.6	70.1	10.9	70.5	9.7	70.6	8.2
Surrey Heath	64.5	2.6	65.0	2.8	66.4	3.1	67.7	3.2	68.3	2.8	69.5	2.4	70.3	1.8	71.0	1.4
Aylesbury Vale	65.4	<0.001	66.2	<0.001	66.7	<0.001	67.2	<0.001	67.2	<0.001	68.1	<0.001	68.6	<0.001	68.9	<0.001
Bracknell and Ascot	64.5	5.6	64.7	5.2	65.5	5.0	65.7	4.4	66.1	4.1	66.5	3.5	66.7	3.0	67.2	2.6
Chiltern	64.9	10.4	65.4	10.6	65.9	11.3	67.0	12.3	67.7	11.3	68.4	9.1	69.2	7.0	70.1	5.4
Newbury and District	65.2	5.7	65.7	5.8	65.8	5.1	66.6	4.7	67.1	4.2	68.1	3.8	68.4	3.3	69.1	2.9
North and West Reading	66.0	5.0	66.0	4.6	66.6	4.2	66.6	3.8	67.6	3.8	67.6	3.0	68.0	2.5	68.3	2.1
Oxfordshire	66.2	25.9	67.1	27.4	67.5	26.7	68.0	26.1	68.2	22.0	69.0	20.2	69.4	16.5	69.9	13.7
Slough	62.3	4.3	62.5	4.1	63.2	4.1	63.3	3.3	64.2	3.1	64.1	2.7	65.0	2.5	65.6	2.0
South Reading	62.5	3.8	62.6	3.6	62.7	3.4	62.7	2.9	63.7	2.7	63.2	2.1	64.2	2.0	64.3	1.6
Windsor, Ascot and Maidenhead	66.1	<0.001	66.2	<0.001	66.3	<0.001	67.0	<0.001	66.7	<0.001	66.5	<0.001	66.5	<0.001	66.7	<0.001
Wokingham	66.6	6.5	67.1	6.0	67.4	6.3	67.8	5.8	68.3	5.0	68.7	4.5	68.9	4.0	69.1	3.5
Dorset	66.9	33.9	67.6	35.8	68.0	36.8	68.4	35.2	68.9	32.1	69.5	26.5	70.1	20.7	70.7	16.0
Fareham and Gosport	63.5	11.5	64.5	11.8	64.9	10.9	65.2	9.7	65.7	8.8	66.2	7.8	66.5	6.7	67.3	6.1
Isle of Wight	63.3	6.3	64.0	6.6	64.7	7.3	65.3	6.8	65.9	6.4	66.8	5.5	67.7	4.2	68.8	3.5
North East Hampshire and Farnham	62.0	<0.001	63.3	<0.001	64.9	<0.001	67.0	<0.001	69.1	<0.001	71.7	<0.001	73.9	<0.001	76.2	<0.001
North Hampshire	63.7	6.4	64.1	6.5	64.8	6.6	65.9	6.9	67.2	7.1	68.0	5.9	69.2	5.0	69.9	3.8
Portsmouth	62.0	7.6	62.5	8.2	63.1	7.6	64.3	8.2	64.9	7.0	65.9	6.0	66.8	4.8	67.6	3.8
South Eastern Hampshire	63.6	8.4	64.2	9.1	64.9	9.9	65.6	9.9	66.3	9.2	67.3	8.0	68.1	6.7	68.9	5.3
Southampton	62.9	<0.001	63.5	<0.001	64.3	<0.001	65.2	<0.001	66.3	<0.001	67.0	<0.001	67.6	<0.001	68.4	<0.001
West Hampshire	66.1	17.2	66.9	18.7	67.4	20.0	68.1	20.5	68.9	18.9	69.8	15.7	70.4	12.0	71.3	9.1

4.7 Discussion

In this first research chapter we aimed to summarise and monitor survival for all cancers combined in England at both national and local level. For this purpose, we designed a summary survival indicator using a three-way standardisation technique that combines, in a weighted average, individual cancer survival estimates for pre-specified combinations of sex, age group and cancer type. The new indicator was named index of cancer survival and was envisioned as a simple tool to aid health policy-makers monitor the effectiveness of cancer care services at both national and local level.

For England, the survival index was estimated at one-, five- and ten-years after diagnosis for selected periods over the 40 years analysed (1971-72, 1980-81, 1990-91, 2000-01, 2005-06 and 2010-11). The CCG index was estimated at one-year after diagnosis for each of the 16 years of diagnosis between 1996-2011. Net survival was the measure chosen to estimate the individual cancer survival components for the indexes. To estimate these net survival components, the first choice of estimator considered was the non-parametric Pohar-Perme estimator that is the current gold standard estimator of net survival. However, the estimation of survival at both national and local level had similar challenges due to the large number of components for which survival estimates were needed, some of them having small number of cases and events leading to very unstable estimates with small precisions and large variances. For the England index, 185 net survival estimates were needed for each combination of period and times after diagnosis, adding to a total of 3,330 net survival estimates. For the CCG index, 35 net survival estimates were needed for each year of diagnosis, adding to a total of 118,160 net survival estimates for all CCGs. Given these challenges, we decided to develop a modelling strategy using excess hazard regression models to improve the estimation of net survival for the components of the indexes. The modelling strategy was developed by setting up a priori 8 candidate models fitted sequentially and retaining those models with the lowest AIC as the best fitting models. All models included age at diagnosis and year of diagnosis as main effects. Age was included in its continuous form in all the models instead of using age-groups to borrow strength across all the ages. During the post-estimation net survival was estimated for each age-group by averaging the survival estimates for all observations with ages falling

within each age-group. Including year of diagnosis in all the models allowed the estimation of more stable trends and improved model fit compared to fitting separate models for each year or period of diagnosis.

The results of the national and the CCG cancer survival indexes have attracted much interest from policy makers and cancer researchers. The results of the England index [1] supported CRUK's vision set out in their 2014 research strategy [155], and they have since been fed into numerous CRUK's public funding campaigns and into online information blogs [156]. The results of the CCG index have been incorporated into annual technical reports published by the Office for National Statistics and Public Health England, which are updated on a yearly basis for the most recent years of cancer diagnosis [2–5]. National policy makers in particular have identified the index of cancer survival as a very useful tool both for national surveillance and for local monitoring of cancer services. As a result, the CCG index was included in the Delivery Dashboard of the NHS' Assurance Framework that sits at the top of NHS accountability tree [157–159] to ensure that local commissioners are held accountable for improving cancer survival in their areas. Cancer researchers worldwide have also been motivated to construct cancer survival indexes for their countries using the three-way standardisation technique we propose. The United States constructed a North American Cancer Survival Index to Measure Progress of Cancer Control Efforts [160] and Japan started to develop a national index of cancer survival (work in progress).

Although the concept of the survival index is simple as it uses a well-known standardisation technique easily understood by non-experts, the estimation of the individual components of the index is long and complex. The modelling strategy we developed for the estimation of the survival components was computationally very intensive, taking from hours up to days for the estimation to be completed due to the large number of models that needed to be fitted. Even so, devising such an approach greatly improved the estimation of net survival, minimising the number of missing estimates for each component of the index and reducing the variance of estimates.

Future research should aim to simplify and optimise the modelling strategy for the cancer survival index. One possibility would be exploring the use of models based on penalised tensor splines recently proposed by Fauvernier et al. [109] for the estimation of cancer

survival. These models have the potential of reducing computational time whilst improving the modelling selection strategy. Another possible alternative to make most of the flexibility of the models used (the use of splines to model the effects of continuous variables) could be based on the following steps: a) predict from the model survival of each combination composed of individual year of age and calendar year, in addition to cancer and sex; b) use the first set of weights, but divide the weights by the number of years included in each age group; c) take the weighted mean of survival estimates to estimate the indexes. Such an approach is derived from a matrix of cancer-year-age-sex combinations rather than directly from the distribution of these combinations in the studied population. It would avoid the issue of missing survival estimates in some categories with sparse data. This approach would rely even more on the robustness of the models as part of the estimates will be out of sample.

In summary, the research presented in this chapter proposes a new monitoring tool for cancer survival at both national and local level. The survival index was designed as a public health measure to assess progress in the overall effectiveness of the health system in diagnosing and managing cancer patients. A novel modelling strategy was developed to improve the estimation of the individual index components. This strategy was presented in a detailed way to facilitate and guide other researchers interested in developing a cancer survival index for their setting. However, as previously mentioned there were several problems regarding the estimation of survival which were not solved with the approach proposed here. In Chapter 6 we aim to address these outstanding estimation challenges by exploring alternative cancer survival models within the Bayesian framework to further improve the estimation of cancer survival. In addition to these estimation challenges, we have also faced difficulties in understanding spatial patterns and trends from the results of the CCG index presented in Tables 4.10 and 4.11. Interpretation was very challenging, almost impossible, due to the large number of estimates unfolding over very large tables. Using 'standard' data visualisation tools, such as ranked bar charts or thematic maps could provide misleading visual interpretations of results when dealing with such a large number of estimates, in particular when mapping estimates for smaller areas. The next research chapter will explore data visualisation techniques for survival outcomes focused on improving the visualisation of both the national and CCG cancer survival indexes.

Chapter 5

Data visualisation techniques for cancer survival relevant to health policy

"... few people will appreciate the music if I just show them the notes. Most of us need to listen to the music to understand how beautiful it is. But often that's how we present statistics; we just show the notes we don't play the music." Prof. Hans Rosling

In this chapter we aimed to improve the visualisation of cancer survival for a more successful dissemination to policy-makers (Research Aim 2). A joint smoothing and mapping technique is adapted to produce smooth small-area cancer survival maps. Funnel plots are extended to visualise the spread of individual cancer survival estimates around a pre-specified target value by formulating the correct control limits for cancer survival. An application of these two techniques is presented to visualise the results of the index of net survival for CCGs (estimated in Aim 1), and to exemplify how the same set of results can be used for national surveillance and local monitoring of cancer survival.

5.1 Introduction

Successful cancer survival studies do not only depend on the statistical ability to produce robust survival estimates, but also on how effectively these findings are communicated to a vast range of audiences, including the research community, the media, the general public and health policy-makers [161]. Effective dissemination of research evidence is important in bridging the gap between research and policy [162]. Geographical variations in cancer survival have commonly been presented using a conventional set of representations [35–37, 163, 164], as for instance:

- Tables of results, listing point estimates of cancer survival with some associated measure of variability (95% confidence intervals, standard errors or precisions), stratified by health geographies, years (or periods) of diagnosis or other factors.
- ‘Bar charts’, ranking survival estimates in ascending or descending order of survival.
- Thematic or choropleth maps, showing spatial patterns by colouring each area according a pre-specified colour scale of survival.

These ‘classical’ ways of presenting cancer survival outcomes have attracted much criticism. Important cancer survival patterns are not easily identified from a long table presenting thousands of survival estimates, such as the results for the CCG cancer survival index presented in Tables 4.10 and 4.11 (Chapter 4). Bar charts ranking survival based on the value of the point estimates disregard any variability associated with those estimates. This can lead to spurious ranking in particular when estimates are based on small number of cases and events. Minor changes in the point estimates can result in big jumps in the ranking due to greater instability of estimates [165]. Thematic maps displaying survival estimates for small areas can be difficult to interpret due to excessive variation (or noise) masking true survival patterns.

5.2 Smoothing technique for small-area cancer survival maps

A small-area based smoothing technique for cancer risk mapping has been developed by colleagues at the Finnish cancer registry since the 1980s [166–168], together with a dedicated software to produce the smoothed maps. This technique has been used to produce several updates of the Cancer Incidence Atlas in Nordic countries [169] and other Cancer Incidence Atlases [170]. As an example, Figure 5.1 illustrates the outcome of the smoothing technique applied to lung cancer incidence for women diagnosed in Finland (these maps were produced by colleagues at the Finnish cancer registry with permission from Prof. Pukkala to be used).

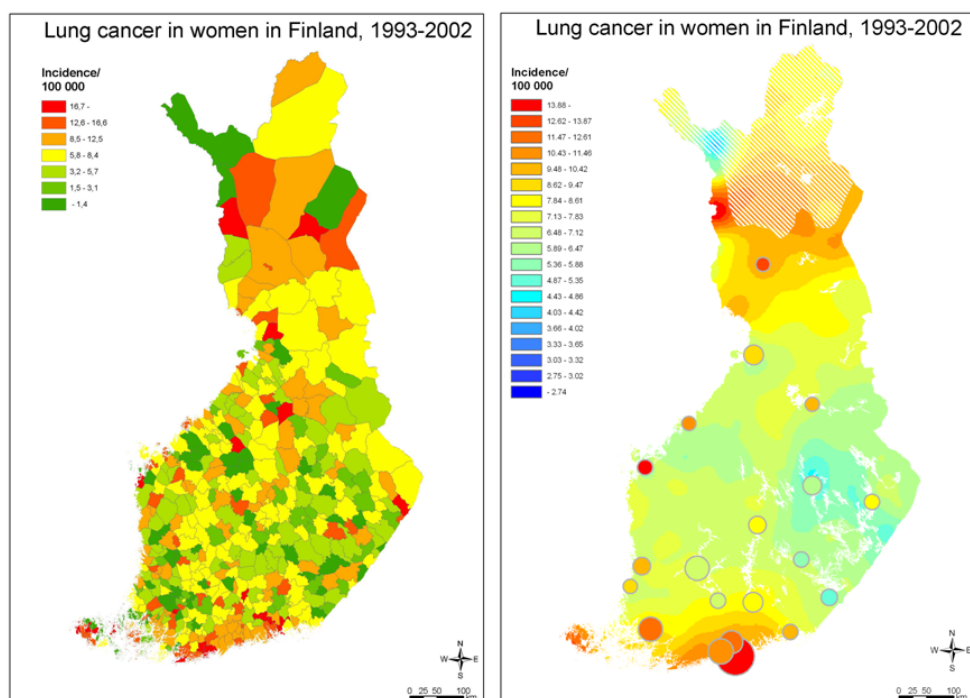


Figure 5.1: Lung cancer incidence for women diagnosed in Finland (with permission to use from Prof. Pukkala at the Finnish cancer registry)

The left panel maps the incidence rates estimated for each small-area separately without filtering out the excessive variation. This makes it difficult to identify any spatial pattern present in the data. The smoothed map on the right panel shows floating weighted averages of incidence after applying the smoothing technique, making it much easier to interpret the emerging spatial pattern. For instance, it allowed to uncover the hotspot of very high incidence around the Finnish capital of Helsinki. The areas in the map are usually irregularly

shaped polygons, defined by their centroids, but the final 'smooth' map will not show the individual area boundaries.

To better visualise the urban-rural contrast in cancer incidence, the smoothing software allows observations for cities with large populations not to be included in the smoothing. Instead, their rates are shown in circles superimposed on the smoothed background. The diameter of each circle is relative to the population size of the city, and the colour of each city circle indicates the lung cancer incidence. The choice of these cities is optional and user-defined.

We have adapted the same joint smoothing and mapping technique to filter out excessive variation from small-area cancer survival maps. This work was done in collaboration with the developers of this technique at the Finnish cancer registry (Professor Eero Pukkala and Mr. Toni Patama) [168]. The software was adapted by the programmer Mr. Tony Patama during several work visits to the Cancer Survival Group at LSHTM. The adapted software was installed for sole use by the Cancer Survival Group.

In order to create a smoothed map of cancer survival, the first step is to estimate a priori a set of small-area specific cancer survival for each area separately using any of the estimators described in chapter 2. These estimates are then supplied to the software to create smoothed maps. The technique implemented in the software can be summarised as follows:

A map covering the areas of interest is uploaded into the software using the respective shape-files of the area boundaries to create a thematic map. This map is then layered on a raster grid, i.e. a gridded array of small cells sized $Y_{km} \times Y_{km}$ (Figure 5.2). The size of the grid cells is a smoothing parameter that needs to be defined in the software to ensure that the final smoothed maps achieve the desired resolution. For instance, if a smoothed map is intended to be displayed on an internet page, the suggested map resolution should be between 300pixels \times 300pixels and 700pixels \times 700pixels. Higher resolutions such as 2048pixels \times 2048pixels (or more) are advised when producing high resolution maps for peer-review publications. Depending on the size of the area of interest, the size of the grid cells will differ to achieve the desired map resolution. The end-user of the smoothing software has the option of tuning these parameters by producing a few different smoothed

maps to find the best map resolution for the desired purpose. As guidance, the software developers suggest that for a country of the size of Finland, a grid cell size of $2\text{km} \times 2\text{km}$ is adequate, but for the Netherlands, a much smaller country geographically, a grid cell size of $500\text{m} \times 500\text{m}$ is advised as sufficient [171].

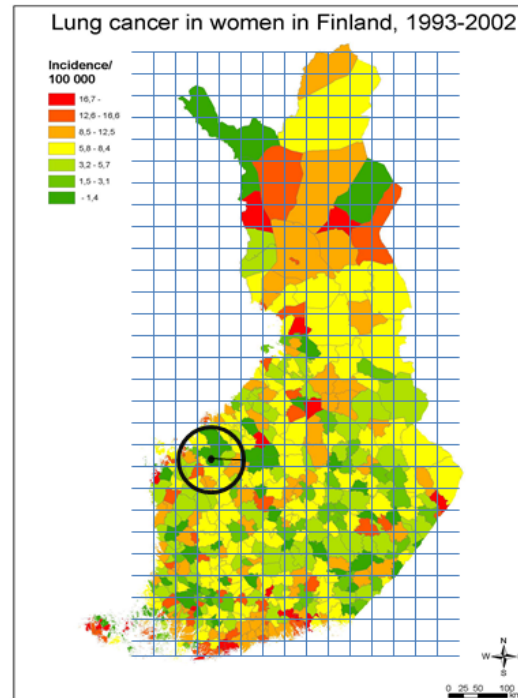


Figure 5.2: Map of lung cancer incidence for women diagnosed in Finland with overlaid raster grid

For each grid cell, a cancer survival value is calculated as a weighted average of all the cancer survival estimates for the areas that fall within a circle of radius r from the middle of each grid cell. Figure 5.3 shows the circular smoothing window centered at a grid point \mathbf{x} with a circle radius r and a distance between the grid cell center and an area centroid d_i .

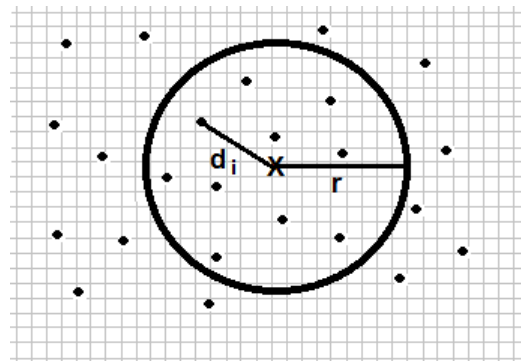


Figure 5.3: Circular window defining distance from center of grid point

The weighted cancer survival average for each grid cell can be written as,

$$CS_j = \frac{\sum_{n=1}^i W_i \cdot CS_i}{\sum_{n=1}^i W_i} \quad (5.1)$$

where CS_j is the weighted cancer survival average for grid j ($j = 1, 2, 3, \dots$), CS_i is the cancer survival estimate for area i ($i = 1, 2, 3, \dots$) and W_i is the corresponding weight for that area.

To calculate the smoothing weights W_i , we start by calculating a weight w_i for each distance d_i of area i . These weights are drawn from a bell-shaped weighting function (Figure 5.4) known as the Butterworth's function [171] that defines the weights as inversely proportional to distance (d_i). The parameter d_0 is the distance at which the weights w_i are halved ($w_i=0.5$). As an example, Figure 5.4 shows a decay function for $d_0=15\text{km}$.

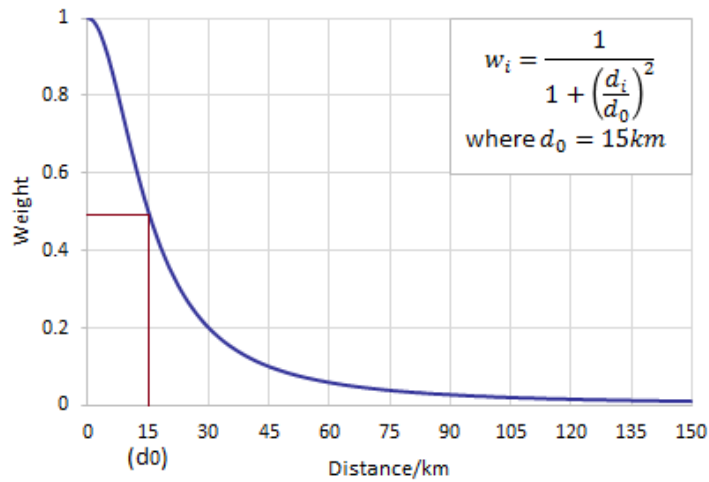


Figure 5.4: Smoothing decay function

As shown in Figure 5.4, the formula to calculate the weights w_i for each distance d_i of area i ($i = 1, 2, 3, \dots$) is written as,

$$w_i = \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \quad (5.2)$$

The final smoothing weights W_i are adjusted using the population size P_i of each area as,

$$W_i = w_i \cdot P_i \quad (5.3)$$

Smoothing parameters

There are two additional smoothing parameters used in the interpolation for which optimal values need to be calculated: 1) the radius r for the circular smoothing window and 2) the distance parameter d_0 that defines the distance weights.

Based on intensive testing and tuning of the interpolation parameters, the software developers advise that the interpolation performs best when the value of the radius r is approximately 10-fold compared to the distance parameter d_0 [171]. They also caution that several iterations might be required to find the optimal value for parameter d_0 . If the distance d_0 is set too short (in km), the cancer survival estimates for some smaller areas can become visible on the map, whereas if d_0 is set too long some areas might lose all variation with a risk of over-smoothing.

As a final step, when the interpolation is completed and a cancer survival value is calculated for each grid cell, a colour will be allocated to each cell based on its cancer survival value to produce the final smoothed map. Whilst the software developers advise that to map cancer incidence (using a relative scale), a 19-colour scale is optimal to provide the map a smoother visual colour transition, for cancer survival (using an absolute scale) a 15-colour scale was sufficient to provide a smooth colour transition.

The next section will introduce the second data visualisation technique proposed in this research chapter to better visualise cancer survival outcomes. Funnel plots are extended to visualise the spread of individual cancer survival estimates around a target value and by formulating the correct control limits for cancer survival outcomes. This work is presented in the form of a tutorial paper. At the end of the chapter an application of these two visualisation techniques (smoothed maps and funnel plots) will be presented using the results of the CCG cancer survival index estimated in Chapter 4.

5.3 Funnel plots for population-based cancer survival [6]

Funnel plots are simple graphical tools designed to detect excessive variation in performance indicators by simple visual inspection of the data whilst avoiding spurious ranking of results as seen with bar charts. They have been used extensively in meta-analysis studies to detect publication bias [172, 173], but more recently they have been recommended as the most appropriate way to display variation in performance indicators for a vast range of health-related outcomes, such as risk-adjusted mortality rates [174, 175].

As illustrated in Figure 5.5 constructing a funnel plot involves plotting a series of estimates on the y-axis (indicator or outcome) against a measure of the precision of these estimates on the x-axis. A target (horizontal line) is then superimposed on the plot representing the desired expectation for the outcome. A set of control limits is drawn around the target acting as cut-off points, beyond which individual estimates are identified as having a divergent behaviour from what is expected given the target. The control limits form a funnel shape around the target, with wider limits for smaller precisions reflecting the larger variability expected for these estimates.

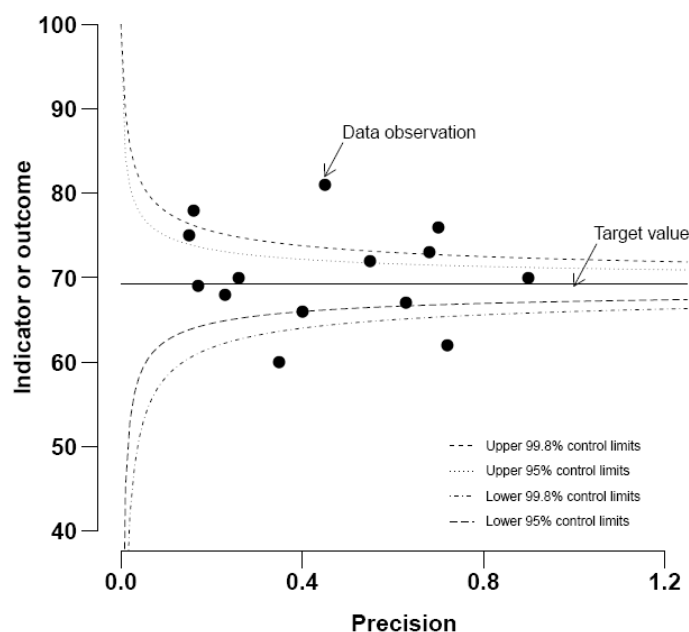


Figure 5.5: Funnel plot illustration

5.3.1 Research publication 2

The second research publication was prepared as a tutorial article extending the use of funnel plots to visualise several cancer survival outcomes. The article provides a step-by-step guide to construct funnel plots and defines the correct formulation of the control limits for cancer survival and excess hazards. It also presents three applications using the different measures to familiarise the reader with the uses and interpretation of these plots.

Example R code to construct a funnel plot is provided in [Appendix A](#).

The article was peer-reviewed and published in *Statistics in Medicine*. The final published article is inserted from next page.

Copyright © 2013 John Wiley & Sons, Ltd. Green Open Access.

Title: 'Funnel plots for population-based cancer survival: principles, methods and applications' [\[6\]](#)

Authors: Manuela Quaresma, Michel P Coleman and Bernard Rachet

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	199304	Title	Ms
First Name(s)	Manuela		
Surname/Family Name	Quaresma		
Thesis Title	Population-based cancer survival at small area level: methodological developments		
Primary Supervisor	Professor Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	2013		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	Published within PhD registration period		
Have you retained the copyright for the work?*	No	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.


SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	MQ derived the formulas for the funnel plots and structured the tutorial paper supervised by BR. MQ wrote the article. MQ, MPC and BR commented on the structure and revised the article.
--	---

SECTION E

Student Signature	
Date	20/11/2019

Supervisor Signature	
Date	20/11/2019

Funnel plots for population-based cancer survival: principles, methods and applications

M. Quaresma,^{*,†} M. P. Coleman and B. Rachet

Funnel plots are graphical tools designed to detect excessive variation in performance indicators by simple visual inspection of the data. Their main use in the biomedical domain so far has been to detect publication bias in meta-analyses, but they have also been recommended as the most appropriate way to display performance indicators for a vast range of health-related outcomes. Here, we extend the use of funnel plots to population-based cancer survival and several related measures. We present three applications to familiarise the reader with their interpretation. We propose funnel plots for various cancer survival measures, as well as age-standardised survival, trends in survival and excess hazard ratios. We describe the components of a funnel plot and the formulae for the construction of the control limits for each of these survival measures. We include three transformations to construct the control limits for the survival function: complementary log-log, logit and logarithmic transformations. We present applications of funnel plots to explore the following: (i) small-area and temporal variation in cancer survival; (ii) racial and geographical variation in cancer survival; and (iii) geographical variation in the excess hazard of death. Funnel plots provide a simple and informative graphical tool to display geographical variation and trend in a range of cancer survival measures. We recommend their use as a routine instrument for cancer survival comparisons, to inform health policy makers in planning and assessing cancer policies. We advocate the use of the complementary log-log or logit transformation to construct the control limits for the survival function. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: population base; cancer survival; funnel plot; surveillance; geographical variation

1. Introduction

Funnel plots are designed to detect excessive variation in performance indicators by simple visual inspection of the data [1]. They borrow their underlying theory from statistical process control [2], which comprises a range of statistical tools developed to monitor manufacturing processes and to ensure their compliance to pre-defined standards with minimum variation. The most popular of these tools is the 'Shewhart control chart' [3], in which repeated measurements of an ongoing process are plotted against time, and three horizontal lines are superimposed: the mean and its upper and lower control limits. The control limits act as thresholds, beyond which a particular estimate is considered to be 'out of control', and the reasons for its divergent behaviour should be investigated. This visual representation of data makes it very easy to detect estimates that lie outside the control limits. Such outsiders indicate either that the process might be subject to greater variability than would be expected from random variation alone or that these estimates are outliers, that is, their true means are very different from the others.

The principle of funnel plots is similar to that of Shewhart control charts, except that the estimates are generally made at a single point or period of time, and they are plotted against a function of their statistical precision, instead of against time. The control limits thus take the shape of a funnel, instead of two lines parallel to the x -axis. The wider control limits to the left give greater emphasis to the increased variability expected from less precise estimates, while the narrower limits to the right emphasise the

Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

*Correspondence to: M. Quaresma, Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

†E-mail: Manuela.Quaresma@lshtm.ac.uk

reduced variability from precise estimates. The main use of funnel plots in the healthcare sector has so far been for meta-analyses, in particular to detect publication bias [4], but they have recently been strongly recommended as the most appropriate way to display performance indicators for a wide range of outcomes, such as comparisons of risk-adjusted rates between healthcare units or surgical outcomes [1, 5]. Funnel plots avoid spurious ranking of the individual estimates, which can arise with ranked bar charts [6]. Because they can be easily extended to control for any measurable outcome [1], it makes them particularly useful for population-based cancer survival, both to display geographical variation and as a surveillance tool in local health authorities.

We have applied funnel plots to four different types of measure associated with population-based cancer survival: (i) relative survival [7]; (ii) age-standardised relative survival [6, 8, 9]; (iii) trends in relative survival [7]; and (iv) risk-adjusted excess hazard ratios (EHRs) [10, 11]; but the formulation of the control limits for each of these measures has not been published. In this article, we present these different measures, and we show that the same methodology can be applied for other estimators of cancer survival.

The components of funnel plots are defined in Section 2.1. In Section 2.2, the survival measures of interest are briefly described. The mathematical expressions for the control limits for each indicator are detailed in Section 2.3. Section 3 illustrates the application of these funnel plots using real data, and Section 4 discusses some limitations of the funnel plots in this context and outlines future methodological developments.

2. Methods

2.1. Components of the funnel plot

A funnel plot comprises four elements [1]: the outcome variable (or indicator), the target (or reference) value for the outcome, a precision parameter, and a set of control limits (Figure 1).

Given a set of estimates of the outcome, as for example a set of 1-year relative survival estimates for 152 small areas, the funnel plot is constructed by plotting these estimates on the y -axis against their associated precision on the x -axis, forming a scatter plot. The precision parameter is a natural choice to represent the statistical accuracy of each estimate, and it can be taken as any function proportional to the inverse of the variance. Other functions can also be used, such as the inverse of the standard error [12].

The target (solid horizontal line, Figure 1) is then superimposed: this is a constant value, considered independent of the observations, and it specifies the expected value for the outcome. It may be taken as either the average of the individual estimates or a single estimate obtained from the pooled data, such as the national average, or any externally chosen value, against which each of the estimates is to be compared.

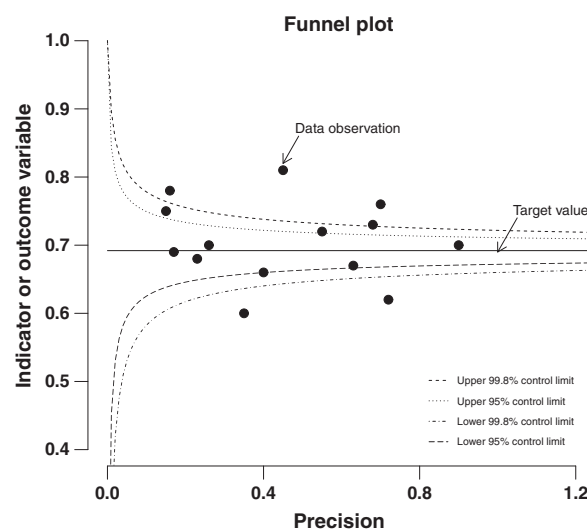


Figure 1. Illustration of the funnel plot components.

The control limits are also independent from the individual estimates. They depend only on the target, and their correct formulation depends on the underlying theoretical distribution of the target value. Control limits (dashed lines, Figure 1) for a given level of significance (α) are drawn around the target across the entire observed range of precision of the individual estimates. The most common levels of significance are $\alpha = 5\%$ and $\alpha = 0.2\%$ so that the resulting 95% and 99.8% control limits represent approximately two and three standard deviations, respectively, from the target value at each level of precision. These thresholds act as alerts or alarms [13, 14]. An estimate that appears outside the control limits is identified as diverging from the specified target and is an out-of-control estimate, or a probable outlier that may need to be investigated further.

2.2. Cancer survival and related measures

2.2.1. Cancer survival. Population-based cancer survival studies aim primarily at estimating the net survival, that is, the survival that the cancer patients would experience if the cancer of interest was the only possible cause of death. Net survival is a theoretical quantity that can be estimated within two main settings: the cause-specific setting and the relative-survival setting. Cause-specific survival requires the cause of death to be known reliably for each cancer patient, which is rarely the case at population level. For this reason, relative survival is usually preferable for population-based analyses because the cause of death is not required; relative survival is conventionally defined as the ratio between the survival observed amongst the cancer patients and the survival that would have been expected had these patients only been subject to the all-cause mortality (by age, sex, calendar period, etc.) in the general population from which they are drawn [15]. The expected (or ‘background’) mortality is obtained from population life tables. Relative survival estimates produced by cancer registries around the world over the last 50 years have generally been based on one of three approaches to deriving the expected survival from the life tables: Ederer-1 [16], Ederer-2 [17] and Hakulinen [18]. However, these approaches lead to biased estimation of net survival if the informative censoring due to differential competing risks is not properly accounted for [19]. A new non-parametric estimator of net survival has recently been proposed that takes informative censoring into account, thus producing unbiased estimates of net survival [19].

Net survival can also be estimated via a modelling approach in which the observed hazard of death is modelled as the sum of two components: the hazard due to the cancer (commonly designated as the excess hazard of death) and the hazard due to all other causes (the background hazard) [20–23]. This approach allows for the estimation of the excess hazard of death, whilst adjusting for the influence of covariables, such as age and stage at diagnosis. Once the excess hazard has been estimated, an estimate for net survival can be easily derived from the mathematical relationship between the hazard and survival functions, as long as the effects of the variables responsible for the informative censoring, such as age, are correctly specified in the model. The most common models are based either on the framework of generalised linear models [22] or on the full maximum likelihood approach [21]. Several extensions have been proposed to enable modelling of the baseline hazard with flexible parametric functions, inclusion of time-dependent covariables to allow for non-proportional effects and the use of flexible functions to account for non-linear excess hazards from continuous covariables. Several packages are available in STATA and R software to estimate net survival using these estimators [23–26].

Different survival quantities can be derived using the aforementioned net survival estimators. The most commonly reported are as follows: estimates of the cumulative survival up to a specific time after diagnosis, say, 5 years after diagnosis; interval-specific survival (e.g. survival between the second and third years after diagnosis); and conditional survival (e.g. survival up to 5 years after diagnosis, conditional on 1-year survival). These survival quantities can, in turn, also be age standardised to take into account differences in the age distribution of cancer patient populations. Age standardisation is crucial when the purpose of the analysis is to compare survival estimates between regions or countries, or over calendar time, so that the comparison is not masked by differences in the age profile of the cancer patients. The most usual age-standardisation procedure is the direct method, a weighted average of the age-specific survival estimates:

$$AS = \sum_{i=1}^n w_i \times S_i$$

where, for a given time since diagnosis, AS is the age-standardised survival and S_i is the survival (for any given survival function as defined earlier) for patients diagnosed in the i^{th} age group, $i = 1, \dots, n$. A common age grouping used in standardisation of survival in adults ($i = 5$) is 15–44, 45–54, 55–64,

65–74 and 75–99 years. w_i is the set of age-specific weights in the standard cancer patient population [27].

Direct standardisation requires that the summation of weights across all age groups is unity. The standard error of the age-standardised survival ($se(AS)$) is a weighted average of the standard errors of the age-specific estimates:

$$se(AS) = \sqrt{\sum_{i=1}^n w_i^2 \times se(S_i)^2}$$

Funnel plots can be constructed for each of the survival quantities described earlier by using the formulations for the control limits provided in this article.

2.2.2. Trends in cancer survival. The measures described so far represent survival information as a snapshot in time or for a given calendar period. Funnel plots can also be constructed to display differential survival trends over time between regions, or between population sub-groups. For this purpose, trends in survival can be quantified by fitting a variance-weighted least squares regression to, say, 5-year survival estimates in successive calendar years (or periods). Depending on how the regression model is specified, regression coefficients can be interpreted as the difference in survival trends between population sub-groups defined by socio-economic deprivation [28, 31] or by area of residence [7]. These coefficients can be used to construct a funnel plot. The regression models to quantify trends in survival can be fitted in most of the common statistical software packages, such as STATA or R.

2.2.3. Excess hazard ratio. Funnel plots can also be constructed to display variation between areas in the EHR. Multivariable regression models can be used to estimate the EHR for each unit of analysis. The EHRs derived from the regression models usually represent the ratio of the excess hazard of death for each unit to that in a chosen reference unit. In a funnel plot, however, we want to avoid any individual estimate being the reference to which all other estimates are compared. We suggest the modification of the regression model to a deviation contrast, so that the estimates for each unit of analysis are compared to the overall or ‘grand’ mean excess hazard. This is an unweighted average of the excess hazard in all categories and is taken as the target. Funnel plots can then be constructed using either the EHRs or the log-EHRs (the coefficients estimated by the model). If the EHR is used, a logarithmic transformation should be applied. The log-EHRs can be assumed to follow an approximately Normal distribution, and the formulae for the control limits are thus given by the usual expression for a Normal interval.

2.3. Formulation of the control limits

We will describe the formulation of the control limits for the three survival metrics (survival estimates, survival trends and EHRs). An asymptotic Normal distribution can be applied to a suitable transformation of the metric to avoid that, at extreme values of event time, the approximate control limits include impossible values outside the range $[0, 1]$. In general terms, considering $g(\tau)$ as a function of the metric τ so that

$$g(\tau) \sim \text{Normal}(g(\theta), 1/\rho_{g(\tau)})$$

where $g(\tau)$ represents the transformed function of the metric τ ; θ represents the target for the metric τ ; and $g(\theta)$ is the transformed target value. For survival, the target will be the overall or average survival. For survival trends, this will be the overall trend for all data combined. For the EHR or its logarithmic transformation, the target will be the corresponding metric derived from the pooled data. $\rho_{g(\tau)}$ is the precision of the transformed function of the metric τ , and $1/\rho_{g(\tau)}$ represents its approximate variance, obtained via the Delta method (Appendix).

The control limits for the transformed function $g(\tau)$ are given by

$$g(\theta) \pm z_{1-\alpha/2} \times \sqrt{1/\rho_{g(\tau)}}$$

where $\pm z_{1-\alpha/2}$ represents the upper and lower percentile limits of the standard Normal distribution ($z = 1.96$ for 95% control limits and 3.09 for 99.8% control limits).

The control limits for the measure τ itself are then obtained by back-transforming the lower and upper control limits estimated for the transformed function.

Table I. Formulae for the control limits using transformations of the survival metrics.

Transformation	Lower control limit	Upper control limit
Identity: $g(\tau) = \tau$	$\theta - z_{1-\frac{\alpha}{2}} \times \sqrt{1/\rho_g(\tau)}$	$\theta + z_{1-\frac{\alpha}{2}} \times \sqrt{1/\rho_g(\tau)}$
Complementary log-log: $g(\tau) = \log(-\log(\tau))$	$\theta^{\exp\left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta \times \log(\theta)}\right)}$	$\theta^{\exp\left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta \times \log(\theta)}\right)}$
Logit: $g(\tau) = \log\left(\frac{\tau}{1-\tau}\right)$	$\frac{1}{1 + \left(\frac{1-\theta}{\theta}\right) \times \exp\left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta \times (1-\theta)}\right)}$	$\frac{1}{1 + \left(\frac{1-\theta}{\theta}\right) \times \exp\left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta \times (1-\theta)}\right)}$
Logarithmic: $g(\tau) = \log(\tau)$	$\theta \times \exp\left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta}\right)$	$\theta \times \exp\left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\rho_g(\tau)}}{\theta}\right)$

Table I provides general formulae for the lower and upper control limits for the measure τ using different transformations.

Survival trends (for example, the average annual change in survival or the slope coefficient) and the log-EHR can be considered as Normally distributed, so the identity transformation is applied, which is equivalent to no transformation. For the survival function, three main transformations have been proposed in the literature: complementary log-log, logit and logarithmic [29, 32]. The complementary log-log and the logit transformations remove the constrain for the limits of the survival function to be within the range of values [0, 1] (see Appendix for more details). Because the logarithmic transformation allows the upper limit to exceed 1 in some situations, we discourage the use of this transformation. In our experience, both complementary log-log and logit transformations behave in a very similar manner (data not shown). In this paper, we have applied the log-log transformation on the survival function.

3. Examples

3.1. Relative survival by a small area in England

One-year relative survival was estimated for all women diagnosed with breast cancer in England during 1996–2006 and followed up to the end of 2007. The survival estimates for each of 152 small areas (Primary Care Trusts (PCT)) covering the whole country were used to construct funnel plots for patients diagnosed in 1996, 1999, 2002 and 2005 (Figure 2) [8]. Each data point represents the estimate for one PCT. The target value is taken as the overall (pooled) national average for England in each year of diagnosis. A funnel plot offers a simple, visual approach to understanding the survival trends (Figure 2). The eight PCTs with survival below the control limits in 1996 (solid black circles) are considered as low ‘outliers’. They are highlighted in all the plots in order to trace whether survival improves in those PCTs or remains consistently low over time. Similarly, the high outliers are identified with solid grey circles.

Both the national average survival and the PCT-specific survival estimates generally increased between 1996 and 2005. In 1996, many PCTs fell outside the 95%, but this ‘over-dispersion’ was much less evident in later years, with a convergence of most estimates towards the target and a higher proportion of PCTs falling inside the control limits, across a wide range of precision.

3.2. Racial and geographical variation in age-standardised relative survival

In the CONCORD study, a worldwide population-based study, racial and geographical variations in cancer survival were displayed in funnel plots for 22 US cancer registries [6]. Figure 3 shows 5-year age-standardised relative survival for women diagnosed with colorectal cancer during 1990–1994 and followed up to the end of 1999. Survival estimates were stratified by race (Black and White people) and by the two federal cancer registration systems: the Surveillance, Epidemiology and End Results (SEER) Program and the National Program of Cancer Registries (NPCR). The target for the funnel plot, at around 0.60, was taken as the pooled age-standardised 5-year relative survival for all participating registries.

The funnel plot shows that colorectal cancer survival among Black people (solid symbols) was consistently lower than among White people (open symbols): survival ranged from 0.45 to 0.57 in Black people, and between 0.54 and 0.66 in White people. Survival for areas covered by SEER (circles) was

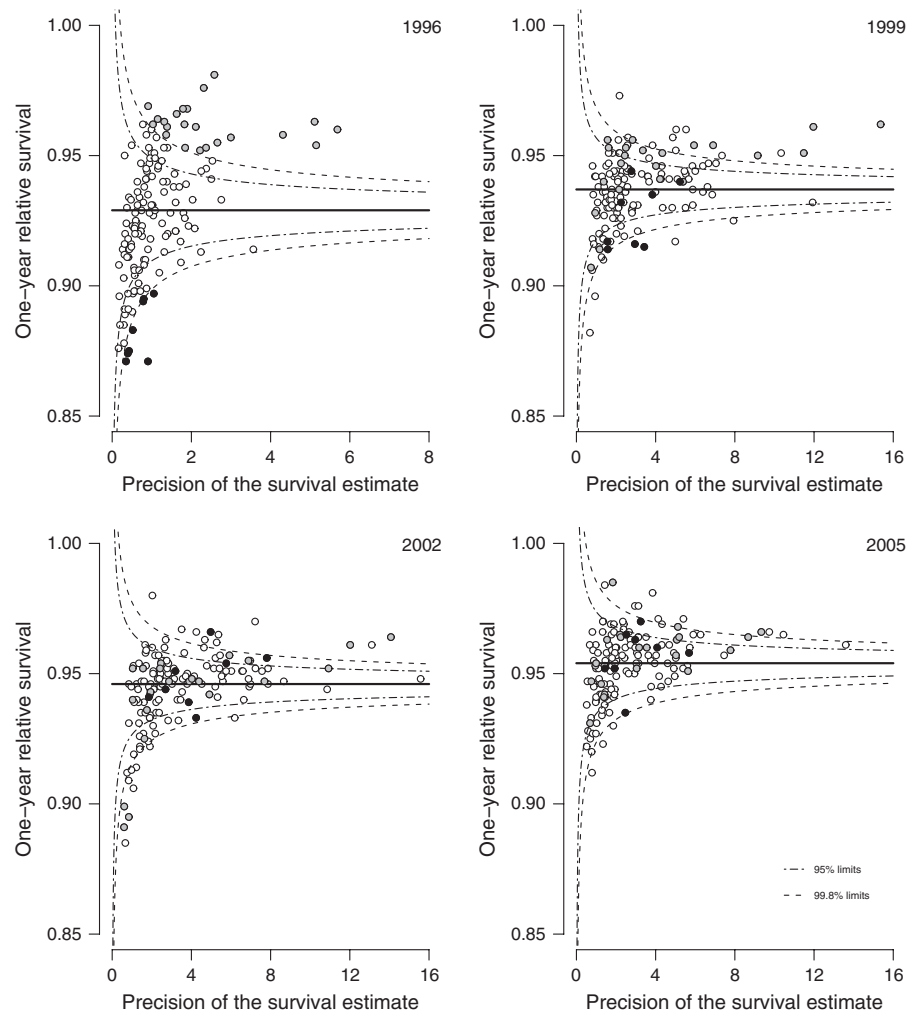


Figure 2. Funnel plots of 1-year relative survival, women (15–99 years) diagnosed with breast cancer in 152 Primary Care Trusts in England: 1996, 1999, 2002 and 2005.

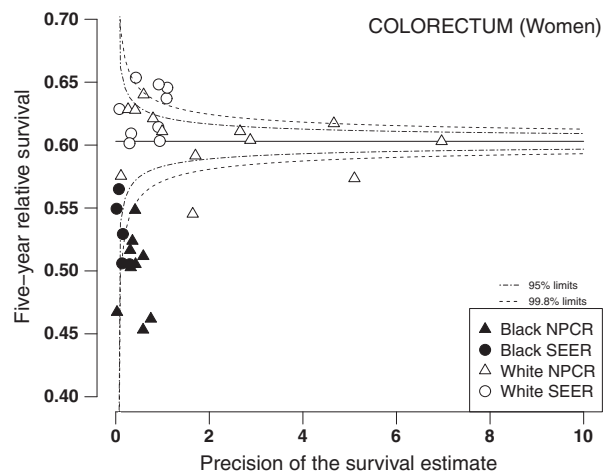


Figure 3. Funnel plot of 5-year age-standardised relative survival, women (15–99 years) diagnosed with cancer of the colon and rectum combined during 1990–1994 and followed up on 31 December 1999, 16 US states and six metropolitan areas, by race and cancer registration system: National Program of Cancer Registries (NPCR) and Surveillance Epidemiology and End Results (SEER).

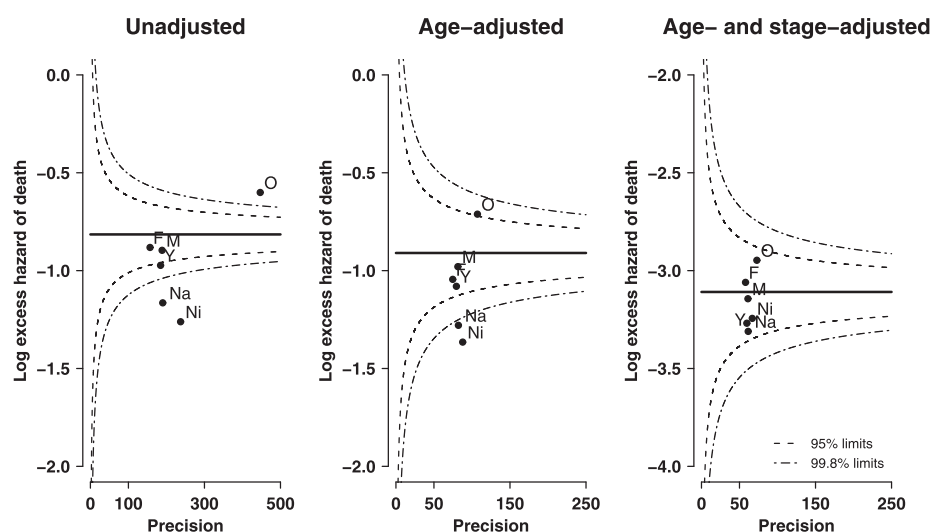


Figure 4. Funnel plots of the unadjusted, age-adjusted, and age-adjusted and stage-adjusted log-excess hazard ratios for deaths within 5 years of diagnosis, by prefecture in Japan (Y, Yamagata; M, Miyagi; Ni, Niigata; O, Osaka; F, Fukui; and Na, Nagasaki): lung cancer (women).

generally higher than in NPCR areas (triangles), for both White and Black people. Black people generally experienced lower survival than in the target value, and when living in NPCR areas, their survival was often below the lower control limits. By contrast, survival among White people was more often within the control limits. Survival for White people was outside the control limits for a few areas, either above, for three SEER areas, or below, for two NPCR areas. This was observed for estimates of both low and high precision.

3.3. Excess hazard ratio

Regional differences in the excess hazard of death from lung cancer in women in six Japanese prefectures were estimated using multivariable Poisson regression models for relative survival [11]. The funnel plots show the log-EHR at 5 years after diagnosis for each prefecture. The raw excess hazards were successively adjusted for age, then for age and tumour stage (Figure 4).

The unadjusted and age-adjusted EHRs were below the lower 95% control limit (lower than the average hazard of death) for Niigata and Nagasaki, but after adjusting for stage at diagnosis, the EHRs were all within the control limits. Similarly, after adjusting for age and stage, the EHR for the prefecture of Osaka, initially an upper outlier with significantly high excess mortality, also fell within the control limits. The unadjusted geographical disparities were thus mostly explained by differences in the stage distribution, with cancer patients in Niigata and Nagasaki generally being diagnosed at an earlier stage than elsewhere.

4. Discussion

This paper extends the use of funnel plots to population-based cancer survival and parameters derived from it and provides the formulae for the control limits for each of these measures. Funnel plots are a simple and informative approach for hospital-based comparisons [1]. We show here that funnel plots can also be used to examine population-based data such as geographical variation in cancer survival and that they represent a valuable tool to inform health policy makers in both planning and evaluating the effectiveness of cancer policy. Estimates that fall outside the upper or lower limits are easily identified as having divergent behaviour. The readability and interpretability of the plots can be greatly improved with the use of different symbols to distinguish between different groups (Figure 3) or by tracing the performance of estimates that are initially outliers over time (Figure 2). Funnel plots do not identify divergent estimates on the basis of their low or high ranking in league tables, which can lead to unnecessary investigations. Funnel plots should not, however, be used as a formal statistical test for multiple comparisons.

Indicators are often displayed in funnel plots without adjustment for case mix or other confounders, but when the data are available, such adjustment is essential for the validity of a funnel plot in detecting outlying performers [30]. Funnel plots of unadjusted estimates of a given parameter presented alongside plots that are successively adjusted for key confounders can offer additional explanatory value (Figure 4) [11]. Such adjustment is crucial to limit over-dispersion, when the majority of the data points fall outside the funnel. Over-dispersion may be due to insufficient risk adjustment of the outcome measure, or it may indicate that the indicator is not the most appropriate [13]. Over-dispersion may also occur when the point estimates are derived from a multivariable model using a large amount of data. For example, in order to obtain more robust survival estimates for a given small area, several years of data may be used to reduce variability of individual estimates. Artificially inflating the variance around the target has also been suggested [13] and has produced satisfactory results (data not shown).

A common limitation on wider use of funnel plots for cancer survival has been the lack of correct specification of the control limits and their unavailability in standard statistical packages. We specify here the control limits for a range of cancer survival estimates. We advocate the use of complementary log–log or logit transformations for the survival function.

Sterne and Egger [12] have published guidelines on the choice of metric to plot on the x -axis for detecting bias in meta-analysis. These authors argue that plotting the outcome (y) against the standard errors (x), rather than the precision, would be desirable because the control limits become straight lines in a form of a funnel instead of two curvy limits. They cite as disadvantages the fact that studies with smaller sample sizes will be compressed at the bottom of the funnel in the presence of very large studies. However, Spiegelhalter [1] suggested that for institutional comparisons, the best choice for funnel plots is a function of the precision, which provides a more natural and direct interpretation of the x -axis. For example, with outcomes that follow a binomial distribution, the number of cases can be plotted on the x -axis as a function of the precision. The precision of the survival function does not have such a convenient interpretation; it is nevertheless far simpler to interpret the impact of precision on the parameter estimates in a funnel plot than with the widely varying confidence intervals in a ranked bar chart.

In conclusion, funnel plots are a simple and powerful tool to display outcomes derived from population-based cancer survival data, and we recommend them as a routine tool for cancer survival comparisons, to improve the planning and evaluation of cancer policies locally, nationally and worldwide.

Appendix

Derivation of the control limits for the survival function

All the transformations discussed in the succeeding paragraphs are made on the original scale of the survival estimates, that is, within the range 0–1. To display the funnel plots for survival on the percentage scale, the multiplication by 100 should be made on the final values, after back-transformation.

Consider $S(t)$ as a survival function with T the survival time: the 100 $(1 - \alpha)$ % control limits for $S(t)$ can be obtained by first applying a transformation to $S(t)$ that will remove the constraint of the values to be in $[0, 1]$ (for the complementary log–log and logit transformation) and transform them into the range $(-\infty, \infty)$. The 100 $(1 - \alpha)$ % control limits are then obtained for the transformed values, which are then back-transformed to the original scale in the range $[0, 1]$.

The variance for the transformed survival function can be estimated using the Delta method. Generically, let X be a random variable and $g(X)$ a function of X . The Delta method is applied by using the first two terms of a Taylor series expansion around the mean of the variable to approximate the value of the function as follows:

$$g(X) \cong g(\mu) + (X - \mu) \times g'(\mu)$$

where $g'(\mu) = \frac{\partial g(X)}{\partial x}|_{X=\mu}$, so that the variance estimator of the function is then approximately given by

$$\text{var}(g(X)) = \text{var}(g(\mu) + (x - \mu) \times g'(\mu)) = g'(\mu)^2 \times \text{var}(x - \mu) = g'(\mu)^2 \times \sigma_x^2$$

1. Complementary log–log transformation

Considering a complementary log–log transformation of $S(t)$ given by $\log(-\log(S(t)))$, with \log denoting the natural logarithm, we have

$$Z_{\log(-\log(S(t)))} = \frac{\log(-\log(\hat{S}(t))) - \log(-\log(S(t)))}{\sqrt{\hat{\text{var}}(\log(-\log(\hat{S}(t))))}} \sim \text{Normal}(0, 1)$$

The 100 (1 - α) % control limits for $\log(-\log(S(t)))$ are given by

$$[\text{lower limit, upper limit}] = \log(-\log(S(t))) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\hat{\text{var}}(\log(-\log(S(t))))}$$

The Delta method can be used to derive the variance estimator for $\log(-\log(S(t)))$ as

$$\text{var}(\log(-\log(S(t)))) = [(\log(-\log(S(t))))']^2 \times \text{var}(S(t)) = \left[\frac{1}{-\log(S(t))}\right]^2 \times \text{var}(S(t))$$

so that

$$[\text{lower limit, upper limit}] = \log(-\log(S(t))) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\text{var}}(S(t))}{(S(t) \times \log(S(t)))^2}}$$

The 100 (1 - α) % control limits for $S(t)$ are then obtained from the limits for the $\log(-\log(S(t)))$ and using the inverse complementary log-log transformation:

$$y = \log(-\log(S(t))) \Rightarrow S(t) = \log(-\log(S(t)))^{-1} = \exp(-\exp(y))$$

so that

$$\begin{aligned} & \log(-\log)^{-1} \left[\log(-\log(S(t))) \mp z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{\hat{\text{var}}(S(t))}}{S(t) \times \log(S(t))} \right] \\ &= \exp \left(-\exp \left(\log(-\log(S(t))) \mp z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{\hat{\text{var}}(S(t))}}{S(t) \times \log(S(t))} \right) \right) \end{aligned}$$

Simplifying the preceding expressions and re-writing the variance term as a function of the precision, that is, $\text{var}(S(t)) = 1/\rho_{S(t)}$, where ρ is the precision, the control limits for the survival function can be written as

$$[\text{lower limit, upper limit}] = \left[S(t)^{\exp \left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t) \times \log(S(t))} \right)}, S(t)^{\exp \left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t) \times \log(S(t))} \right)} \right]$$

2. Logit transformation

Considering a logit transformation of $S(t)$ given by $\log(\frac{S(t)}{1-S(t)})$ so that

$$Z_{\text{logit}(S(t))} = \frac{\text{logit}(\hat{S}(t)) - \text{logit}(S(t))}{\sqrt{\hat{\text{var}}(\text{logit}(S(t))))} \sim \text{Normal}(0, 1)$$

The 100 (1 - α) % control limits for $\text{logit}(S(t))$ are given by

$$[\text{lower limit, upper limit}] = \text{logit}(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\hat{\text{var}}(\text{logit}(S(t)))}$$

Using the Delta method to derive the variance estimator for the $\text{logit}(S(t))$, we obtain

$$\text{var}(\text{logit}(S(t))) = \text{var}(\log(\frac{S(t)}{1-S(t)})) = [(\log(\frac{S(t)}{1-S(t)}))']^2 \times \text{var}(S(t)) = \frac{1}{(S(t) \times (1-S(t)))^2} \times \text{var}(S(t))$$

so that

$$[\text{lower limit, upper limit}] = \log(\frac{S(t)}{1-S(t)}) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\text{var}}(S(t))}{(S(t) \times (1-S(t)))^2}}$$

The 100 (1 - α) % control limits for $S(t)$ are then obtained from the limits for the $\text{logit}(S(t))$ and using the inverse logit transformation:

$$y = \text{logit}(S(t)) \Rightarrow S(t) = \text{logit}(S(t))^{-1} = \frac{1}{1 + \exp(-y)}$$

$$\text{so that } \text{logit}^{-1} \left[\text{logit}(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\text{var}}(S(t))}{(S(t) \times (1-S(t)))^2}} \right]$$

$$= \frac{1}{1 + \exp \left(-\log(\frac{S(t)}{1-S(t)}) \pm z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{\hat{\text{var}}(S(t))}}{S(t) \times (1-S(t))} \right)}$$

Simplifying these expressions and re-writing the variance in terms of the precision lead to the following control limits for the survival function:

$$\begin{aligned} [\text{lower limit, upper limit}] &= \left[\frac{S(t)}{S(t) + (1-S(t)) \exp \left(z_{1-\frac{\alpha}{2}} \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t)(1-S(t))} \right)}, \frac{S(t)}{S(t) + \frac{S(t)}{(1-S(t))} \exp \left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t) \times (1-S(t))} \right)} \right] \\ &= \left[\frac{1}{1 + \left(\frac{1-S(t)}{S(t)} \right) \exp \left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t) \times (1-S(t))} \right)}, \frac{1}{1 + \left(\frac{1-S(t)}{S(t)} \right) \exp \left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t) \times (1-S(t))} \right)} \right] \end{aligned}$$

3. Logarithmic transformation

Considering a logarithmic transformation of $S(t)$ given by the natural $\log(S(t))$ so that

$$Z_{\log(S(t))} = \frac{\log(\hat{S}(t)) - \log(S(t))}{\sqrt{\hat{\text{var}}(\log(S(t)))}} \sim \text{Normal}(0, 1)$$

The 100 (1 - α) % control limits for $\log(S(t))$ are given by

$$[\text{lower limit}, \text{upper limit}] = \log(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\hat{\text{var}}(\log(S(t)))}$$

Using the Delta method to derive the variance estimator for $\log(S(t))$ as

$$\text{var}(\log(S(t))) = [(\log(S(t)))']^2 \times \text{var}(S(t)) = \frac{1}{S(t)^2} \times \text{var}(S(t))$$

$$\text{so that } [\text{lower limit}, \text{upper limit}] = \log(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\text{var}}(S(t))}{S(t)^2}}$$

The $100(1 - \alpha)\%$ control limits for $S(t)$ are then obtained from the limits for the $\log(S(t))$ and using the inverse log transformation:

$$y = \log(S(t)) \Rightarrow S(t) = \log(S(t))^{-1} = \exp(y)$$

$$\text{so that } \log^{-1} \left[\log(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{\hat{\text{var}}(S(t))}}{S(t)} \right]$$

$$= \exp(\log(S(t)) \mp z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{\hat{\text{var}}(S(t))}}{S(t)})$$

Simplifying the expressions and re-writing the variance in terms of the precision give the following control limits for the survival function:

$$\begin{aligned} [\text{lower limit}, \text{upper limit}] &= \left[\frac{S(t)}{\exp(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t)})}, S(t) \times \exp(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t)}) \right] \\ &= \left[S(t) \times \exp\left(-z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t)}\right), S(t) \times \exp\left(z_{1-\frac{\alpha}{2}} \times \frac{\sqrt{1/\hat{\rho}_{S(t)}}}{S(t)}\right) \right] \end{aligned}$$

Acknowledgement

We would like to thank Dr. Yuri Ito, researcher at the Department of Cancer Epidemiology and Prevention at the Osaka Medical Center for Cancer and Cardiovascular Diseases in Japan, for providing the data for the construction of the funnel plots for the EHR example.

References

1. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; **24**:1185–1202.
2. Shewhart WA. Economic Control of Quality of Manufactured Product. *New York: D. Van Nostrand Company* 1931:501.
3. Shewhart WA. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal American Statistical Association* 1925; **20**:546–548.
4. Egger M, Smith D, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**:629–634.
5. Mayer EK, Bottle A, Rao C, Darzi AW, Athanasiou T. Funnel plots and their emerging application in surgery. *Annals of Surgery* 2009; **249**:376–383.
6. Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T, Micheli A, Sant M, Weir HK, Elwood JM, Tsukuma H, Koifman S, Azevedo e Silva G, Francisci S, Santaquilani M, Verdecchia A, Storm HH, Young JL, the CONCORD Working Group. Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncology* 2008; **9**:730–756.
7. Rachet B, Maringe C, Nur U, Quaresma M, Shah A, Woods LM, Ellis L, Walters S, Forman D, Steward J, Coleman MP. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncology* 2009; **10**:351–369.
8. Quaresma M, Jakomis N, Gordon E, Carrigan C, Coleman MP, Rachet B. Index of Cancer Survival for Primary Care Trusts in England—Patients Diagnosed 1996–2009 and followed Up to 2010. *Office for National Statistics*, 2011.
9. Walters S, Quaresma M, Coleman MP, Gordon E, Forman D, Rachet B. Geographical variation in cancer survival in England, 1991–2006: an analysis by Cancer Network. *Journal of Epidemiology and Community Health* 2011; **65**:1044–1052.
10. Coleman MP, Rachet B, Quaresma M, Lepage C, Baum M, Sikora K. Bradford NHS Trust and Panorama. *Lancet* 2006; **368**:730–731.
11. Ito Y, Ioka A, Tsukuma H, Ajiki W, Sugimoto T, Rachet B, Coleman MP. Regional differences in population-based cancer survival between six prefectures in Japan: application of relative survival models with funnel plots. *Cancer Science* 2009; **100**:1306–1311.
12. Sterne J, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001; **54**:1046–1055.
13. Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Quality Safety Health Care* 2005; **14**:347–351.
14. Spiegelhalter DJ. Funnel plots for institutional comparison. *Quality Safety Health Care* 2002; **11**(4):390–391.
15. Berkson J, Gage RP. Calculation of survival rates for cancer. *Proceedings of the Staff Meetings of the Mayo Clinic* 1950; **25**:270–286.
16. Ederer F, Axtell LM, Cutler SJ. The relative survival: a statistical methodology. *National Cancer Institute Monograph* 1961; **6**:101–121.

17. Ederer F, Heise H. Instructions to IBM 650 Programmers in Processing Survival Computations. *Methodological note No. 10, End Results Evaluation Section, National Cancer Institute, Bethesda MD*, 1959.
18. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982; **38**:933–942.
19. Perme MP, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012; **68**(1):113–120.
20. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Journal of the Royal Statistical Society: Series C* 1987; **36**:309–317.
21. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; **9**:529–538.
22. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* 2004; **23**:51–64.
23. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal* 2010; **9**:265–290.
24. Cancer Research UK Cancer Survival Group (2006). strel computer program, version 1.2.8. Downloaded from <http://www.lshtm.ac.uk/eph/ncde/cancersurvival/tools/> on the 1st February 2009. Faculty of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, UK.
25. Dickman PW, Coviello E, Hills M. *Estimating and modelling relative survival in SAS and Stata*, 2006. Downloaded from www.pauldickman.com/rsmodel/index.php on the 11th June 2009, version 1.2.8.
26. Pohar M, Stare J. Relative survival analysis in R. *Computer Methods and Programs in Biomedicine* 2006; **81**(3):272–278.
27. Corazzari I, Quinn MJ, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer* 2004; **40**:2307–2316.
28. Coleman MP, Babb P, Damiecki P, Grosclaude PC, Honjo S, Jones J, Knerer G, Pitard A, Quinn MJ, Sloggett A, Stavola BLD. Cancer survival trends in England and Wales 1971–1995: deprivation and NHS Region. Studies on Medical and Population Subjects. *The Stationery Office* 1999; **61**.
29. Anderson JR, Bernstein L, Pike MC. Approximate confidence intervals for probabilities of survival and quantiles in life-table analysis. *Biometrics* 1982; **38**(2):407–416.
30. Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *British Medical Journal* 2003; **326**(7393):786–788.
31. Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, Steward J, Brenner H, Estève J, Sullivan R, Coleman MP. Cancer survival in England and Wales at the end of the 20th century. *British Journal of Cancer* 2008; **99** (Suppl. 1):2–10.
32. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics, 2002. John Wiley & Sons, Inc., Hoboken, New Jersey.

5.4 Application: Smoothed maps and funnel plots to visualise the index of cancer survival for CCGs

In this application we demonstrate the use of the two proposed visualisation techniques, smoothed maps and funnel plots, to present the results of the one-year CCG cancer survival index estimated in Chapter 4. The index estimates from Tables 4.10 and 4.11 were used to exemplify how the same set of results can be used for national surveillance and local monitoring of cancer survival. For the purpose of this illustration only results for four selected years of diagnosis were used: 1996, 2001, 2006 and 2011.

5.4.1 Smoothed maps

The smoothing software [171] described in section 5.2 was used to create smoothed maps of England for each year of diagnosis. After tuning the smoothing parameters in the mapping software, the best grid cell size for the interpolation was chosen as 1km x 1km. The best radius for the circular window was 150km and 15km for the distance parameter d_0 .

The CCG boundaries are not shown on the maps as a result of the smoothing process. The open circles on the maps show the survival for CCGs with large populations for which survival estimates are statistically stable and were not included in the smoothing process. For the purpose of this demonstration, we chose four CCGs that included the cities of Liverpool, Sheffield, Birmingham and Coventry. This is an option of the mapping software that allows a selection of such areas to be user-specified as described in section 5.2. In addition, the capital London is shown separately from the main England map for a better view of the results because it is a small area but densely populated.

A 15-colour scale was chosen to provide the smoothest transition in the maps surface appearances. The median of the grid-specific cancer survival was set as the middle point of the scale, with blue shades representing areas with highest survival through to red shades representing areas with lowest survival.

Figure 5.6 presents smoothed maps of England for each year of diagnosis. The smoothed maps show a substantial increase in net survival for England between 1996 and 2011 with estimates ranging between 60-70%. A clear North-South survival gradient can be observed for England and a North-East/South-West gradient for London, with a deficit in survival in the North of England and North-East of London. Despite the overall improvements in survival, the observed disparities are persistent over the years, although slightly reduced.

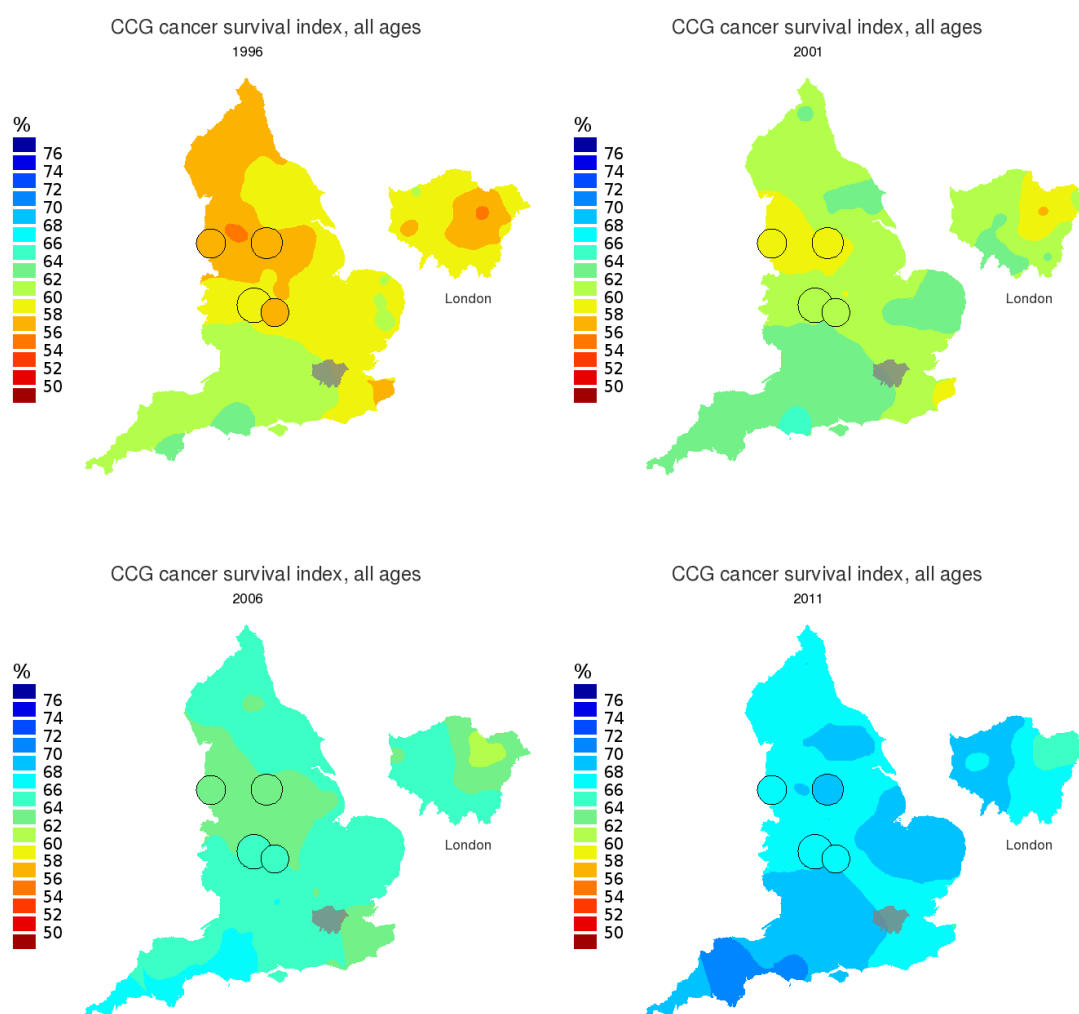


Figure 5.6: Smoothed maps of England using the one-year net survival index for CCGs

5.4.2 Funnel plots

Funnel plots were created using a custom R code we wrote to construct funnel plots for survival outcomes. This R code implements the funnel plot formulation for the control limits using the complementary log-log transformation as defined in research publication 2 [6]. An example code to create one funnel plot can be found in Appendix A.

Each data point in the funnel plots (Figure 5.7) is the estimated net survival index for each of the 211 CCGs. The target was estimated as the mean of all CCG index estimates in each year of diagnosis. The precision values for each CCG index were calculated as the inverse of the variance (Chapter 4). Two sets of control limits were plotted at 95% and 99.8% around the target. CCGs falling below the lower control limit in the 1996 funnel plot (i.e. lower ‘outliers’) are marked in red. These lower CCGs are traced using the same red points in the subsequent funnel plots for 2001, 2006 and 2011. Individual CCGs can be located in the funnel plots using as coordinates their index estimate and precision read from Tables 4.10 and 4.11.

The funnel plots show the spread of the individual CCGs survival indexes around the target value. Overall, one-year net survival increased from around 59% to 68% between 1996-2011 (target value). Survival also increased for all the individual CCGs, with a narrowing of the initial over-dispersion observed in 1996 (i.e. many CCG estimates falling outside of the control limits). Several CCGs which were lower ‘outliers’ in 1996 (red points) improved their survival level and converged within the limits in more recent years, whereas others seem to have worsened (black points below lower limits after 1996).

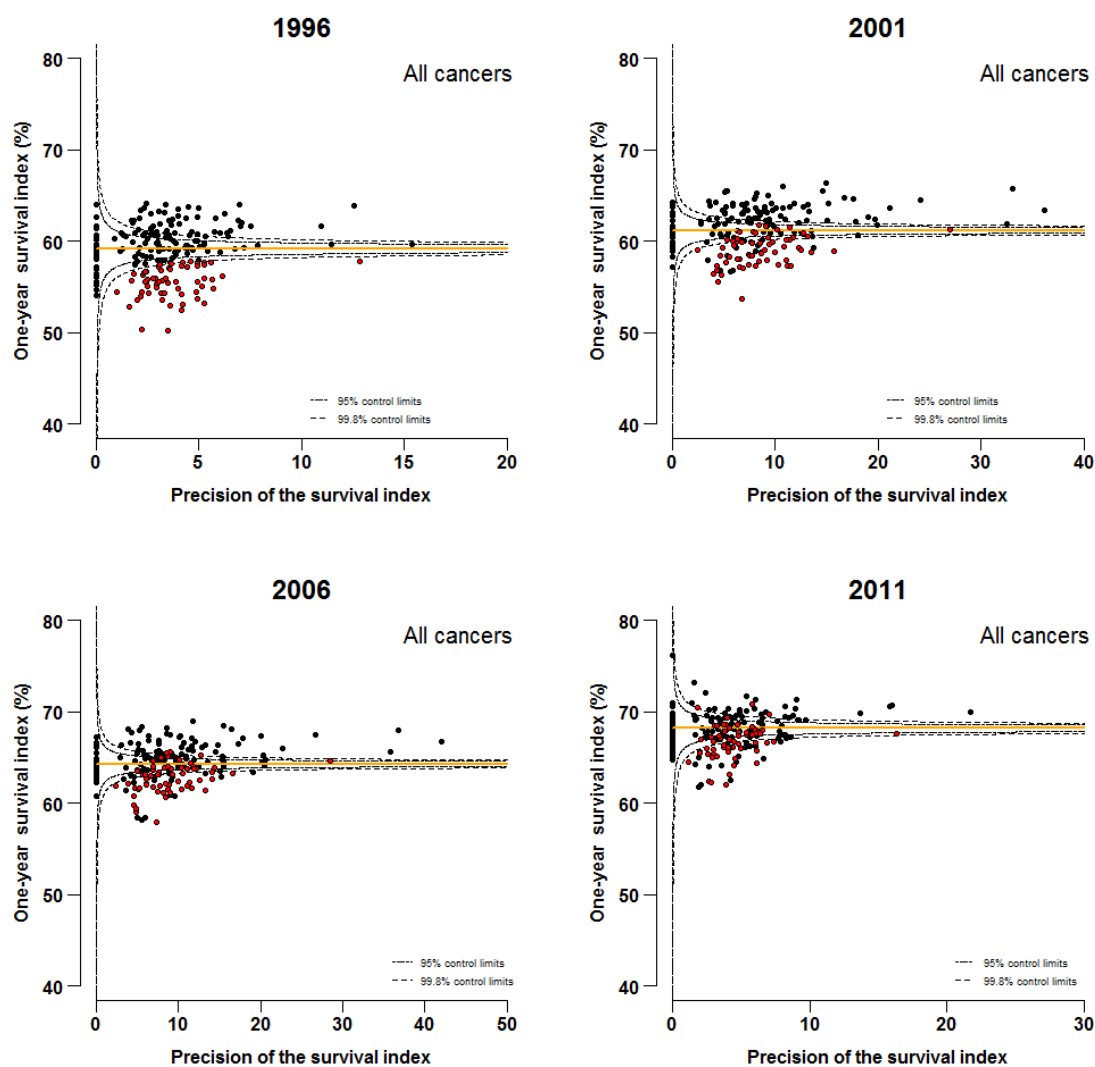


Figure 5.7: Funnel plots of the one-year net survival index for CCGs

5.5 Discussion

In this second chapter we aimed to improve the visualisation of cancer survival for a more successful dissemination to policy-makers. For this purpose we adapted two data visualisation techniques to cancer survival outcomes. First, we adapted a joint smoothing and mapping technique that produces smooth maps based on small-area survival estimates. Next, we have extended funnel plots to visualise the spread of individual survival estimates around a pre-specified target value by formulating the correct control limits for cancer survival outcomes.

To illustrate these two techniques, we have used the results of the CCG cancer survival index estimated in Chapter 4. Smoothed maps provided a 'bird's-eye' view of the cancer survival patterns across the country, after elimination of random local fluctuations present in 'classical' thematic maps. The vivid 15-colour scale provided a smooth and colourful map surface that is more likely to draw attention when presented to a non-expert audience. One limitation of this type of 'surface' smoothing is that it does not take into account the spatial correlation of the data, and thus does not produce standard errors in addition to the estimated smoothed surfaces.

Funnel plots provided a simple yet powerful tool for visualising the spread of survival for individual CCGs whilst avoiding spurious ranking as seen with ranked bar charts. The presence of a target value in the funnel plot (the index estimate for England) discourages direct comparisons between individual CCGs, favouring the comparison of each area against the average level of survival in England.

An additional improvement funnel plots offer over other representations is the easy visualisation of the range of precisions associated with the individual outcome estimates. For instance, in the two long tables of results presented in Chapter 4 (Tables 4.10 and 4.11), identifying any pattern in the range of precisions associated with each CCG survival index was challenging.

As shown by the CCG survival index results, funnel plots are designed to easily detect the existence of over-dispersion in outcomes indicators, that is the presence of greater variability in outcomes than would be expected given the target. Several techniques have been

suggested in the literature to handle over-dispersion in performance indicators [176]. Additional work has been prepared, separate to the research objectives of this thesis, but as an extension of the work presented here to provide a set of guidelines to handle over-dispersion in cancer survival outcomes. The manuscript is currently in review: '*Handling over-dispersion and large precision range in funnel plots for $[0, 1]$ -bounded health-measure estimates*'. Authors: Bannon F. (Queen's University Belfast School of Medicine Dentistry and Biomedical Sciences) and Quaresma M. (LSHTM).

Since publication, smoothed maps and funnel plots have been used by national policy-makers, stakeholders and local cancer managers as a routine monitoring tool [177]. Whilst national policy-makers have used smoothed maps as a strategic planning tool to introduce and update cancer control strategies, funnel plots have helped local cancer managers identify areas with unexpected cancer survival levels, i.e. either much lower or higher than expected compared to the national cancer survival level. Identifying such divergent areas will allow researchers and stakeholders to work together to investigate the reasons associated with such divergent outcomes.

In summary, the research presented in this chapter introduces two data visualisation techniques for cancer survival outcomes. We have demonstrated that smoothed maps and funnel plots are two convenient, effective and 'fair' ways to present cancer survival outcomes both for national surveillance and for local monitoring of cancer survival. As shown, the smoothing technique relies on a 'two-step' process were: first, small-area cancer survival has to be estimated for each area separately, and second, these estimates are then used in the interpolation to produce smoothed maps. However, when in the presence of sparse data, the smoothing technique will not solve the problem of missing survival estimates as we have observed with the CCG cancer survival index estimation. In the next Chapter, we aim to address the outstanding estimation challenges by exploring alternative cancer survival models within the Bayesian framework to further improve the estimation of cancer survival.

Chapter 6

Bayesian approaches for the estimation of cancer survival at small area level

*"Time present and time past.
Are both perhaps present in time future,
And time future contained in time past.
If all time is eternally present
All time is unredeemable..."*

Four Quartets by T. S. Eliot

In this chapter we aimed to determine how Bayesian approaches can be used in the relative survival setting to improve the estimation of cancer survival in the presence of sparse data and when using more complex data structures (Research Aim 3). We started by summarising the existing literature on small area estimation. We then propose a flexible Bayesian excess hazard model formulated on the log-excess hazard scale, and demonstrate how net survival can be estimated from such a model. We demonstrate the applicability of this model by investigating variation in net survival for patients diagnosed with colon cancer living and receiving care in London.

6.1 Introduction

The combined research for Aims 1 and 2 (presented in [chapter 4](#) and [chapter 5](#), respectively) proposed a two-step approach to investigate geographical patterns and time trends in an index of cancer survival defined for small health geographies. An application of the index was presented for the 211 CCGs in England using data for all patients diagnosed with cancer between 1996 and 2011.

In the first step, a modelling strategy was implemented to estimate the individual components needed to construct the index. Separate excess hazard regression models were set-up for each CCG, sex and cancer type, all including age and year of diagnosis. The results showed wide variability in the estimates, with an average of 17% of precision estimates close to zero over the whole period of diagnosis analysed. Between 5-10% of the estimates needed for the construction of the index for each CCG could not be estimated, requiring either model adjustment or replacement of those missing estimates with the estimate obtained for a merged age group of the missing age group with an adjacent (non-missing) age group.

In the second step, two data visualisation techniques were proposed in order to improve interpretability when investigating patterns and trends in the index of cancer survival for both national and local monitoring. Smoothed maps enabled a clearer visualisation of the national patterns of survival in England by using a simultaneous smoothing and mapping technique to filter out excessive variation from less precise estimates. Funnel plots provided an accessible way of displaying the individual index estimates, essential for local monitoring, taking into account the increased variability expected from less precise estimates by defining control limits around a defined target.

The proposed two-step approach substantially improved the investigation of short-term geographical patterns and temporal trends in cancer survival, even when defined at a smaller geographical level, such as CCGs. Despite these improvements, concerns remained regarding the estimation of the 5-10% of 'sex-age-cancer' specific survival components that could not be estimated for the CCG index. In addition, the index was only estimated at one-year since diagnosis because the estimation at later follow-up times was not feasible, due to

the many model non-convergence problems, resulting in very large proportions of missing estimates, that can not be solved by applying the smoothing technique.

The ‘follow-up time’ dimension present in survival analysis adds complexity to the estimation process when compared to incidence or mortality outcomes. For smaller datasets or in the presence of sparse data (mostly due to rare events), the set of patients at risk of dying at any given time since diagnosis is continuously being depleted with the death or censoring of patients. This can lead to unstable estimates with low precision and large variation, even making the estimation of survival not feasible as was observed in our application to CCGs. In addition, when analysing areas of unequal ‘size’, the overall assessment of the geographical patterns can be compromised if the sparse data problem is not taken into account properly in the estimation process. This can result in the real underlying geographical patterns being masked by the presence of too much variation contributed by those more unstable estimates.

The research presented in this Chapter will focus on alternative methods to the regression models used for the estimation of the cancer survival indexes in Chapter 4 to improve the estimation of small-area cancer survival. To the best of our knowledge, the current literature on statistical methods for small-area estimation is fairly limited for cancer survival. We start by presenting an overview of the literature for small-area estimation methods, including quantities other than cancer survival.

6.2 Overview of small-area estimation methods

Small-area estimation (SAE) methodology has been extensively developed in the field of sample surveys [178, 179]. The term ‘small-area’ commonly indicates an area for which the outcome of interest is rare, and does not necessarily refer to the actual size of any given area. For example, a highly densely populated area can be denoted as ‘small’, if the outcome for a rare cancer is the main interest. In this setting, the aim is to obtain estimates of adequate precision when the sample size is not large enough to provide an estimate based solely on the data collected, i.e. a *direct estimate*. Many SAE estimators were developed to provide estimates for non-sampled areas, based on information available for areas that have been sampled and on auxiliary covariates measured for those areas [180, 181], i.e. *indirect* or *model-based* estimates. The most widely used model is the Fay-Herriot model [182], which in its original formulation is defined as a general linear mixed model with area-specific random effects; assumed to be independent, identically and Normally distributed random variables. This formulation makes use of the information available in all the sampled areas, to improve the estimation of quantities for the non-sampled areas. It does not discriminate or make use of information coming only from areas that share common population characteristics with the non-sampled areas, such as might happen with areas that are close to each other or share a common border. More recent developments of SAE models include a spatial dependency structure, enabling the use of information only from selected areas. The main idea behind the Fay-Herriot and other SAE models is to ‘borrow strength’ from data available in sampled areas to help in the estimation of quantities for the non-sampled areas. Most SAE models were developed based on a multilevel model formulation [183, 184], which due to its flexibility allows for different types of effects to be easily modelled; estimation is usually carried out under a frequentist or Bayesian framework.

In epidemiological research, the idea of ‘borrowing strength’ to improve estimation has been at the basis of small area methodology. Commonly, each geography was analysed separately, assuming that data observed for one area were independent from the data observed for another area. Spatial models came to respond to the need of enlarging the set of classic analytical methods, subject to the hypothesis of independence of the observations, to the case of spatial data, where it became evident that the data closer in space have the

tendency to be more similar than those farther away. This ability to incorporate the spatial interdependency, or correlation, of observations makes these models special and able to be applied in a vast range of real situations [185–188].

Spatial studies can be broadly categorised into three types:

1. *Disease mapping*, where the objective is to model and describe the overall spatial distributions or patterns for the disease outcome of interest [189–194].
2. *Spatial correlation or ecological studies*, where the objective is to model the relationship between the spatial distribution for the disease outcome of interest and a group of covariates of interest, which might themselves also present a spatial distribution [195–198].
3. *Disease clustering*, which is concerned with identifying or confirming the existence of unusually high areas of disease in a map. In the case of cancer incidence, the interest would be to identify spots of unusually high incidence, as for example, around a nuclear power station. When looking at survival it would be of more interest to identify spots which present lower survival [199–203].

Spatial data consists of recording the characteristic of interest together with the location at which this characteristic occurred or was measured. For example, a variable might be measured at fixed point locations, giving rise to what is known as geostatistical data. Spatial data can also be measured at locations that are spatial areas, usually called lattice data. A special case of lattice data, are point data, where the exact location at which the observations occur are themselves the variable of interest.

When the exact location (e.g. latitude and longitude) of each observation is known, geostatistical methods [204–207] can be used to estimate and predict, i.e. interpolate continuous risk surfaces for the whole study area and produce isopleth maps. This type of ‘surface’ smoothing can also be done using distance weighting techniques such as moving weighted averages or kernel smoothers [208–210], but geostatistical models have the advantage of taking into account the spatial correlation of the data, and therefore produce standard errors in addition to the estimated risk surfaces. A few studies [211–213] explored the use

of these models for disease mapping, based on small counts and suggest some advantages over the models used for disease mapping (described in the following paragraphs), such as better model flexibility in incorporating the spatial structure, fast computational performance and smaller variance estimates, but applications with real data of geostatistical models to small-area disease mapping, and in particular survival data, are still very scarce [214, 215], and only apply if the aim is to produce a continuous map surface.

In epidemiological studies, lattice data are the most common type of data analysed, usually consisting of counts of cases or sets of observations that occurred in an area of regular or irregular shape. The lattices are referenced by a structure defining the neighbours of each area. This can be done by creating an adjacency matrix, defined based on the Euclidean distance between areas or on whether the areas share a common border.

Seminal work by Besag [216] on Markov Random Fields theory was determinant to the development of many spatial models based on lattices. In his work, Besag has proposed a sub-class of Spatial Markov Random Fields, also termed as auto-models or conditional auto regressive models (CAR), in which the full conditional distributions [217, 218] for the observations in each area can be specified based only on their dependency with their neighbouring areas.

One of the first disease mapping models was proposed to address a specific problem for crude standardised mortality rates (SMRs). As discussed by several authors [204, 219–221], the small counts of cases observed in each of the small areas result in SMRs with extra-Poisson variation, i.e. with variance higher than expected. In order to address this issue, Clayton and Kaldor [219] used a CAR model to accommodate the spatial location of the data in the formulation of an Empirical Bayes [222] model to spatially smooth the relative risks. With this model, unstable estimates can be ‘shrunk’ towards a global or local mean. Different spatial Empirical Bayes models were later proposed [192, 223]. Though very popular, this approach does not consider the extra variance from the estimation of the model parameters, since the estimation of the parameters for the prior distribution is based on the likelihood of the data.

Besag, York and Mollié (BYM) [224] have proposed a full Hierarchical Bayesian model based on a generalised linear mixed model (GLMM) [225], a class of models that resulted

from including random effect terms in Generalised Linear Models (GLM). Such models consist of a fixed effects part related to covariates and a random effect term that depends on the parameters to estimate. The BYM model includes two mutually independent random effect terms that are both spatially and non-spatially structured. The prior distribution for the spatially structured random effect is based on a CAR model and prior distributions are also specified for the parameters of the CAR prior, the so called 'hyperparameters', which are the variance parameters specified at the third level of the hierarchy.

Bayesian models make use of the likelihood of the data and combine it with prior information about the parameters of interest to draw inferences based on the posterior distribution [226, 227]. The prior distribution relates to the distribution of relative risks across the areas studied, so that the neighbour information about the relative risks can be conveniently incorporated in the prior distribution of the parameters. In Bayesian inference, it is usually not possible to obtain the posterior distributions in closed form and numerical integration approaches, such as Laplace approximation and other simulation techniques such as Markov Chain Monte Carlo methods (MCMC), need to be used. Inferences are mostly based on MCMC and involve estimating the quantities of interest from the posterior distribution by drawing samples from the posterior distribution without having to know its closed form [228, 229].

Lawson et al. [230] performed an empirical evaluation based on simulations to analyse the performance of several disease mapping models including non-parametric smoothing methods, empirical Bayes methods and full Bayes methods. Smoothing functions, such as the Nadaraya-Watson kernel smoother [209], were used to filter out the excessive random noise of a map when the estimates for each area were obtained in an univariate way, i.e. without taking into account the spatial dependency of the data. The authors concluded that smoothing methods generally performed poorly whilst the BYM model presented the most robust results.

In a comparison of Bayesian spatial models for disease mapping, Best et al. [231] provide practical guidance on the choice of the prior distribution for the second level of the hierarchy that defines the spatial dependency between the areas. The simulation results show that the BYM model has in general one of the best performances and continues to be the most

chosen model for prior specification. The authors also note that other models can, in some situations, be a better choice to model spatial dependency and that these options should always be explored when conducting a disease mapping study.

The models described until now are all based solely on the specification of the spatial dependency of the data. It can also be useful to look at the evolution of these spatial patterns over time using space-time models. Such models add a temporal effect and a spatio-temporal interaction [232–238]. Other authors proposed joint modelling several diseases sharing common risk or prognostic factors, so that information can be ‘borrowed’ not only in space, but also from other diseases [239–241]. The BYM model formulation was later extended for the joint modelling of two [242] and six [243] diseases.

The approaches described so far relate to the geographical distribution of counts in small areas and they are mainly used for incidence and mortality studies. The concept behind the extension to spatial survival models is similar. Bayesian inferences for survival data have been proposed by several authors, including the use of spatially structured random term effects (frailties) in a hierarchical formulation of spatially correlated survival data, both for geostatistical and lattice data [244–247].

In the field of population-based cancer survival, most of the studies looking at geographical variation estimate survival for each area independently [47, 248, 249]. Very few studies have explicitly included the spatial structure of the data in the model formulation. Exceptions include, Osnes and Aalen [250], who were the first to propose a full hierarchical Bayesian model to smooth cancer survival estimates based on a neighbouring structure. Yu et al. [251] use an Empirical Bayes model to shrink the relative survival estimates towards the global mean but without taking into account the neighbouring structure of the regions. Three other applications have adapted the disease mapping BYM model for the estimation of relative survival [252–254], using the CAR model as a prior distribution. All authors refer to general improvements in the estimation of cancer survival when taking into account the spatial structure of the data compared to univariate models in which the estimation is stratified by area.

Current inference practice for the excess hazard models mentioned above are mainly based on the frequentist maximisation of the likelihood function [100] and very few options are

available for inferences within the Bayesian framework: Fairley et al. [253] proposed a model to examine spatial variation in prostate cancer survival using Bayesian relative survival smoothing within a Generalised Linear Model formulation. The number of events was assumed to follow a Poisson distribution and two random effects were included in the model: a spatially structured random effect for local smoothing and an unstructured random effect global smoothing; Hennerfeind [255] proposed a Bayesian geosadditive relative survival model using penalized P-splines to model the log-baseline effect as well as the nonlinear and time-varying effects of covariates. Spatial and normal random effects were also included in the model formulation; Cramb et al. [256] introduced a Bayesian flexible parametric model which extends a frequentist flexible parametric model on the log cumulative excess hazard scale using restricted cubic splines [114] by adding spatially structured random effects.

In summary, small-area estimation is a common problem in many epidemiological studies and there is a vast body of literature dedicated to address the challenges associated with the instability of estimates. Although, several methods have been proposed, all authors agree on two points: 1) sparse data can compromise the interpretation of important geographical patterns; 2) improved estimation procedures should be used to 'filter' out the excessive variation. Most applications found in the literature are in the context of incidence and mortality studies, with very few applications for cancer survival, reinforcing the need to improve the estimation of cancer survival at the small-area level in order to better understand the origins and mechanisms underlying the observed geographical disparities.

6.3 Flexible Bayesian excess hazard models

The small-area estimation overview presented in the last section, highlighted the need to expand the portfolio of estimation options for small-area population-based cancer survival. Whilst there is a vast amount of different estimation approaches available for incidence and mortality outcomes, methods for overall survival are much less available in general. Population-based cancer survival methodology in particular occupies a very small niche within the survival methodology, and we found very few models addressing the problem of estimation in the presence of sparse or spatially arranged data. The ultimate aim of the small-area estimation overview was to understand what estimation methods are available in the literature, and how the same ideas and principles could be adapted into the relative survival setting. The idea of ‘borrowing’ strength by incorporating the spatial dependency of observations in the model formulation was the main idea shared by the existing modelling approaches. There was also a general consensus that the Bayesian framework lends itself as the most natural choice to implement such models due to its flexibility in incorporating complex data structures, as for instance, using a hierarchical data structure with random effects at the second level of the hierarchy to define the spatial dependency of the areas.

Population-based cancer survival research has seen an active acceleration in methodological developments during the last decade. Several improvements have been proposed to excess hazard regression models, with particular focus on modelling non-linear effects using flexible functions, such as splines, and the correct estimation of net survival at the population level (please refer to the methods overview in Chapter 2). The need for these improvements in estimation methods was partially linked to the increased availability of more complex datasets for population-based cancer research, including more detailed clinical information and new sets of small-area health-geographies. Current inference practice in the cancer survival research community is almost exclusively made within the frequentist framework.

For the purpose of the research presented in this chapter, and considering the reasoning in the last two paragraphs, our ultimate goal is to develop a versatile Bayesian excess hazard model that can be used for small-area estimation by including an adequate dependency structure of the observations and also retain the most recent advances in excess hazard modelling, such as the use of flexible functions to model non-linear effects of covariates.

In order to achieve this ultimate goal, the first step (Aim 3, objective 2) aimed to translate a flexible log-excess hazard model (see Chapter 2, section 2.4.2.2) into the Bayesian framework, since no such model was currently available in the Bayesian literature. The purpose was to first develop a Bayesian counterpart to the ‘basic’ excess hazard model that could later be extended to accommodate ‘random effects’ and other more complex data structures, and to implement a Bayesian post-estimation procedure to derive population-based net survival from such a model. We write ‘random effects’ in quotes because all effects (parameters) are considered random within the Bayesian framework.

This work was prepared as a publication that was peer-reviewed and published [257] in *Statistical Methods in Medical Research*. The article starts by formulating the excess hazard model on the log-scale and proposes the use of low-rank thin plate splines to model the baseline log-excess hazard and the smooth effect of continuous variables. After some consideration, we chose this special type of splines to add flexibility to the model, because the log-likelihood function retains tractability so that numerical integration is not required. This is an important aspect of the proposed model that simplifies the computational burden without sacrificing model flexibility. Another important and innovative aspect this article adds is a step-by-step algorithm to derive posterior distributions of net survival and excess hazards. The model is illustrated with an application to data from patients diagnosed with colon cancer during 2009 in London. We chose colon cancer for this application because we understand well its behaviour from previous experience, including the expected shapes of the excess hazard and the net survival functions for the English population. In addition, as part of extensive checking during the implementation phase of the flexible Bayesian excess hazard model, we have also compared the level of net survival and the excess hazard function with the estimates from the non-parametric Pohar-Perme estimator [91], obtaining very similar results (results not shown).

Complete details can be found in the author accepted manuscript inserted at the end of this section (Research publication 3). In addition, we provide an example R code in Appendix A to implement this model, including the set-up for the low-rank thin plate splines and the post-estimation of net survival.

The next research objective (Aim 3, objective 3) aimed to demonstrate how the flexible Bayesian excess hazard model proposed in the previous work (Aim 3, objective 2) can be extended to model more complex data structures, as a step forward towards our ultimate goal for a versatile excess hazard model accommodating different data structures, including a spatial structure.

We started by demonstrating the practical applicability of extending the flexible Bayesian excess hazard to investigate variation in net survival for patients diagnosed with colon cancer by incorporating a pair of random effects within a hierarchical data structure for patients living within London CCGs and being treated within hospitals in London. A manuscript (Research Publication 4) was prepared based on this work. The full manuscript is inserted from next page and is ready to be submitted to *The Lancet*. The article starts by investigating patterns of patient pathways between the area of residence and the hospital of cancer care. For this purpose, flow maps of London were created to visualise the most frequent pathways between CCGs and hospitals. The variability in cancer survival is then investigated at both CCG and hospital level, after adjusting for some patient and tumour characteristics, such as age at diagnosis, deprivation and stage at diagnosis. To accommodate the hierarchical structure of the data (i.e. that patients within a given CCG of residence or hospital of cancer care are likely to share some characteristics), the flexible Bayesian excess hazard model proposed in the previous section is extended to include a pair of random effects for CCG and hospital. Several innovative graphical representations are used in this work. In addition to the flow maps, windrose graphs arranged according to the approximate cardinal directions of CCGs and hospitals are used to better visualise the proportion of patients by deprivation category and stage at diagnosis. Funnel plots proposed in Chapter 5 are used to display the variability in net survival by CCG and hospital of care.

6.4 Research publication 3

Title: 'Flexible Bayesian excess hazard models using low-rank thin plate splines'.

Authors: Manuela Quaresma, James Carpenter and Bernard Rachet.

Peer-reviewed and published in *Statistical Methods in Medical Research*. Author accepted manuscript inserted from next page.

Copyright © 2019, © SAGE Publications. Green Open Access.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	199304	Title	Ms
First Name(s)	Manuela		
Surname/Family Name	Quaresma		
Thesis Title	Population-based cancer survival at small area level: methodological developments		
Primary Supervisor	Professor Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Statistical Methods in Medical Research		
When was the work published?	2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	Published within PhD registration period		
Have you retained the copyright for the work?*	No	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	MQ developed the Bayesian model supervised by JC and BR. MQ wrote the article. MQ, JC and BR commented on the structure and revised the article.
---	--

SECTION E

Student Signature	
Date	20/11/2019

Supervisor Signature	
Date	20/11/2019

Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate splines

Journal Title
XX(X):1–22
©The Author(s) 2019
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Manuela Quaresma¹, James Carpenter^{1,2} and Bernard Rachet¹

Abstract

Excess hazard models became the preferred modelling tool in population-based cancer survival research. In this setting, the model is commonly formulated as the additive decomposition of the overall hazard into two components: the excess hazard due to the cancer of interest and the population hazard due to all other causes of death. We introduce a flexible Bayesian regression model for the log-excess hazard where the baseline log-excess hazard and any non-linear effects of covariates are modelled using Low Rank Thin Plate splines. Using this type of splines will ensure that the log-likelihood function retains tractability not requiring numerical integration. We demonstrate how to derive posterior distributions for the excess hazard and for net survival, a population-level measure of cancer survival that can be derived from excess hazard models. We illustrate the proposed model using survival data for patients diagnosed with colon cancer during 2009 in London, England.

Keywords

Population-based, cancer, survival, excess hazard, Bayesian, flexible, low rank thin plate splines

Introduction

Regression models for the excess hazard became the preferred modelling tool for cancer survival research using population-based data^{1–3}. In the absence of reliable recording of the cause of death for each cancer patient, these models conveniently allow to filter out the hazard due to other causes of death, whilst focusing inferences on the excess hazard only due to the cancer of interest. In this setting, the model

¹London School of Hygiene & Tropical Medicine, Faculty of Epidemiology & Population Health, London, UK. ²London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, UK.

Corresponding author:

Manuela Quaresma, London School of Hygiene & Tropical Medicine, Faculty of Epidemiology & Population Health, Keppel Street, WC1E7HT London, UK
Email: Manuela.Quaresma@lshtm.ac.uk

is formulated as the additive decomposition of the total hazard into two components: the hazard due to the cancer (the main quantity of interest, also designated as the excess hazard), and the hazard due to all other causes of death, derived from population life tables (also known as background mortality or expected hazard). This set-up allows inequalities in cancer survival to be investigated by looking at the effect of multiple prognostic factors on the form of the excess hazard function or by deriving excess hazard ratios for different sets of characteristics of the population. Model parameter estimates can also be used to derive net survival, both at the individual-level and at the population-level, measuring the survival that can be attributed only to the cancer of interest after accounting for all other causes of death⁴. Net survival estimates can therefore be compared even if the expected hazard differs widely between sub-populations of patients⁵. In their seminal paper, Estève et al.¹ introduced the first regression model for the excess hazard based on the full-likelihood specification using individual survival time data. In its original formulation, the model was proposed on the log-excess hazard scale with the baseline log-excess hazard modelled as a piecewise constant step function, and allowing proportional effects of covariates and linear effects for continuous variables to be modelled. Proposed extensions to this model mainly relaxed the non-proportionality and non-linearity assumptions for covariates and interaction terms, and non-linearity for the baseline excess hazard, by modelling these terms with highly flexible functions such as the commonly used B-splines or restricted cubic splines^{2,6}. The tradeoff for the increased model flexibility obtained with the use of splines, is the added complexity to the likelihood function, that requires advanced numerical integration techniques such as the Cavalieri-Simpson integration² or the Gaussian quadrature⁶ to evaluate the cumulative hazard integral, which will no longer be a tractable function with a closed-form solution.

This applies regardless of the framework of inference, whether frequentist or Bayesian, although inferences for excess hazard models have mainly been based on the frequentist maximisation of the likelihood function. Very few options are available for inferences within the Bayesian framework⁷⁻¹⁰, in particular none describing the process of deriving a posterior distribution of net survival.

The purpose of this article is to introduce a flexible Bayesian regression model for the log-excess hazard, based on individual-level data, with the following characteristics: a) the baseline log-excess hazard is modelled using a flexible function; b) the log-likelihood function retains tractability so that numerical integration is not required; c) the model can accommodate a variety of covariate effects: linear and non-linear (also modelled using a flexible function), proportional and non-proportional; d) one can derive a posterior distribution for the excess hazard, excess hazard ratios and net survival; e) the model can be easily extended to include random effects and hierarchical data structures; f) inference can be done within the Bayesian framework; g) and the model can be implemented using most Bayesian open-source software.

Section 2 specifies the likelihood for the log-excess hazard model, introduces the formulation of the flexible functions used in this article, and describes the Bayesian inference procedure, including the steps to obtain a Bayesian posterior distribution for the excess hazard function, excess hazard ratios and net survival. Section 3 provides an example of application of the proposed model based on the survival time data of patients diagnosed with colon cancer during 2009 in London. Section 4 presents some concluding remarks, discusses the limitations of our study, and proposes further extensions to this work.

Methods

Likelihood formulation for the Excess Hazard Model (EHM)

Let $(t_i, \mathbf{x}_i, \delta_i)$, $i=1, \dots, n$, $t_i > 0$, denote a set of n time to event observations, measured from the date of diagnosis of a cancer until the occurrence of death, with covariates \mathbf{x}_i and vital status indicator δ_i ($\delta_i=0$ if censored, $\delta_i=1$ if death occurred). The likelihood function of the full vector of parameters of interest θ can be written in generic terms as

$$L(\theta) = \prod_{i=1}^n h(t_i, \mathbf{x}_i, \theta)^{\delta_i} \cdot S(t_i, \mathbf{x}_i, \theta) \quad (1)$$

where $h(t_i, \mathbf{x}_i, \theta)$ is the hazard function and $S(t_i, \mathbf{x}_i, \theta)$ is the survivor function. Delayed entry or left-truncation of observations, can be accommodated in the likelihood by including an additional term, $S(t_d, \mathbf{x}_i, \theta)$, representing the survivor function for a pre-specified truncation time $t_d \geq 0$, as

$$L(\theta) = \prod_{i=1}^n \frac{h(t_i, \mathbf{x}_i, \theta)^{\delta_i} \cdot S(t_i, \mathbf{x}_i, \theta)}{S(t_d, \mathbf{x}_i, \theta)} \quad (2)$$

If $t_d > 0$ then $S(t_d, \mathbf{x}_i, \theta) \neq 1$, and the likelihood in equation (2) allows delayed entry of observations²², enabling study designs such as period analysis to be incorporated into the framework²³. If $t_d = 0$ then $S(t_d, \mathbf{x}_i, \theta) = 1$, and the likelihood assumes no delayed entry, simplifying to equation (1). For the purpose of this article, the likelihood in equation (1) is used from here onwards, assuming no delayed entry of observations, but the likelihood in equation (2) could be used equivalently in what follows below.

Considering only the individual contribution of observation t_i , the log-likelihood can be written as

$$\log L(\theta) = \delta_i \cdot \log(h(t_i, \mathbf{x}_i, \theta)) + \log(S(t_i, \mathbf{x}_i, \theta)) \quad (3)$$

Using the following relationship between the survival function and the cumulative hazard function ($H(t_i, \mathbf{x}_i, \theta)$)¹¹:

$$\log(S(t_i, \mathbf{x}_i, \theta)) = -H(t_i, \mathbf{x}_i, \theta) = - \int_0^{t_i} h(u, \mathbf{x}_i, \theta) du \quad (4)$$

and replacing equation (4) into equation (3), the contribution of observation t_i to the log-likelihood can be rearranged as

$$\log L(\theta) = \delta_i \cdot \log(h(t_i, \mathbf{x}_i, \theta)) - \int_0^{t_i} h(u, \mathbf{x}_i, \theta) du \quad (5)$$

An excess hazard model assumes the additive decomposition of the overall hazard, $h(t_i, \mathbf{x}_i, \theta)$, into two components:

$$h(t_i, \mathbf{x}_i, \theta) = h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i) \quad (6)$$

where, $h_E(t_i, \mathbf{x}_i, \theta)$ is the excess hazard function due to the cancer of interest for an observation t_i with \mathbf{x}_i a vector of observed covariates and θ a set of parameters. The second component, $h_P(a_i + t_i, \mathbf{z}_i)$, is the general population hazard function for an observation t_i , evaluated at the attained age at death (or age at censoring): $a_i + t_i$, with a_i the age at diagnosis and \mathbf{z}_i ($\mathbf{z}_i \in \mathbf{x}_i$) a subvector of covariates for which the

population hazard is defined. The population hazard, also known as background mortality, represents the hazard due to all other causes of death than the cancer of interest. It is assumed to be a known quantity, taken as the age-specific mortality rates from existing population life tables, stratified as finely as possible according to a subset of covariates \mathbf{z}_i . This subset of covariates usually contains less covariates than the complete set of covariates available for the cohort of cancer patients, possibly including, in addition to age at death (or censoring), gender and calendar year, socio-economic status, ethnicity or region of residence¹².

Replacing equation (6) into equation (5), the log-likelihood for an excess hazard model can be written entirely as a function of the excess hazard and the population hazard:

$$\log L(\theta) = \delta_i \cdot \log[h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i)] - \int_0^{t_i} h_E(u, \mathbf{x}_i, \theta) du - \int_0^{a_i + t_i} h_P(u, \mathbf{z}_i) du \quad (7)$$

Given that the population hazard $h_P(a_i + t_i, \mathbf{z}_i)$ is a constant, the last integral in equation (7) does not depend on any parameters and thus can be dropped from the log-likelihood, which can be rewritten (up to this constant) as:

$$\log L(\theta) \propto \delta_i \cdot \log[h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i)] - \int_0^{t_i} h_E(u, \mathbf{x}_i, \theta) du \quad (8)$$

Modelling the Log-Excess Hazard function

Equation (8) specifies the log-likelihood for a generic excess hazard model. Inferences can be made by specifying an appropriate model for the excess hazard function ($h_E(t, \mathbf{x})$), which we here assume to have a multiplicative effect of the covariates on the baseline excess hazard. It can be written as,

$$h_E(t, \mathbf{x}) = h_{E_0}(t) \cdot \exp(\beta \cdot \mathbf{x}) \quad (9)$$

where, $h_{E_0}(t)$ is the baseline excess hazard; and $\mathbf{x} = (x_1, x_2, x_3, \dots)$ a vector of observed covariates and $\beta = (\beta_1, \beta_2, \beta_3, \dots)$ the vector of their corresponding parameters. In this article, we propose a model for the logarithm of the excess hazard function, that can accommodate several types of covariate effects. Taking the logarithm of equation (9), we can write, in generic terms, a model for the logarithm of the excess hazard as

$$\log(h_E(t, \mathbf{x})) = \log(h_{E_0}(t)) + \beta_1 \cdot x_1 + g_1(x_2) + g_2(t) \cdot x_3 + \dots \quad (10)$$

where, $\log(h_{E_0}(t))$ is now the baseline log-excess hazard function; β_1 is a linear and proportional effect on the log-excess hazard of covariate x_1 ; $g_1(x_2)$ is a non-linear and proportional effect of a continuous covariate x_2 ; $g_2(t)$ is a non-proportional (i.e. time-dependent) effect of a covariate x_3 .

We choose different constructs of low-rank thin-plate (LRTP) splines to model the baseline log-excess hazard any non-linear effects, and to accommodate time-dependent effects. These first-order polynomials are a penalised type of radial basis splines¹³, that have been discussed by several authors for their simple yet flexible nature, providing a good alternative to other spline constructs, such as B-splines and truncated basis splines. In particular, LRTP splines exhibit fast Markov Chain Monte Carlo (MCMC) convergence properties and conveniently result in tractable likelihood functions^{13,14}. Murray et al.¹⁵ have introduced a

unified framework for flexible, fully Bayesian analysis of overall survival using LRTP splines, providing a detailed description of their formulation (¹⁵: Appendix-A), and also making available user-friendly code for easy practical implementation. We follow this spline implementation in the work presented here, and for completeness, in the next three sections, we provide a brief enunciation of the LRTP splines we use to model the different components of the excess hazard model, but do not go into detail about their implementation.

Modelling the baseline log-excess hazard We start by specifying a partition of the follow-up time range as $0 = \tilde{t}_0 < \tilde{t}_1 < \dots < \tilde{t}_K = \infty$, and following the model formulation published in Murray et al. ¹⁵, we define the model for the baseline log-excess hazard as,

$$\log(h_{E_0}(t; \alpha^*)) = \alpha_0^* + \alpha_1^* t + \sum_{k=2}^K \alpha_k^* (|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \quad (11)$$

where $\alpha^* = (\alpha_0^*, \alpha_1^*, \dots, \alpha_K^*)'$ is the set of spline parameters. Under equation (11), the cumulative excess hazard takes the expression,

$$H_{E_0}(t; \alpha^*) = \sum_{k=1}^K \frac{h_{E_0}(s_k; \alpha^*) [1 - e^{-(s_k - \tilde{t}_{k-1})(u'_{k,K} \cdot \alpha_{(-1)}^*)}]}{u'_{k,K} \cdot \alpha_{(-1)}^*} \quad (12)$$

where $s_k = \max(\min(t, \tilde{t}_k), \tilde{t}_{k-1})$, $\alpha_{(-1)}^* = (\alpha_1^*, \dots, \alpha_K^*)'$, $u'_{k,K} = (1'_k, -1'_{K-k})$, for $k = 1, \dots, K$ with $1'_k$ a k -dimensional vector of ones. Implementation of this spline involves a series of transformations to the spline parameters α^* , as well as constructing a time design matrix and a penalty transformation matrix, as detailed by Crainiceanu et al. ¹⁴ and Murray et al. ¹⁵, so that the baseline log-excess hazard can be rewritten in terms of these transformed components.

Modelling a non-linear effect of a continuous covariate We model any non-linear effect of a generic continuous covariate x , as a smooth effect using a cubic LRTP spline defined as,

$$g(x; \beta^*) = \beta_1^* (x - \bar{x}) + \sum_{j=2}^J \beta_j^* (|x - \tilde{x}_{j-1}|^3 - |\bar{x} - \tilde{x}_{j-1}|^3) \quad (13)$$

where, $\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_J^*)$ is a set of spline parameters, \bar{x} is the sample mean of covariate x , and $(\tilde{x}_1 < \tilde{x}_2 < \dots < \tilde{x}_J)$ is a partition of the covariate's support range. Similarly to the model specification for the baseline log-excess hazard, implementation will require one-to-one transformations to reparametrise (13) in terms of β^* ¹⁵.

Incorporating a non-proportional (time-dependent) effect of a covariate To incorporate a time-dependent effect of a generic covariate x , we use the same time partition as used for the baseline log-excess hazard as in equation (11), and define,

$$\log(h_E(t|x; \alpha^*)) = (\alpha_{0,0}^* + \alpha_{1,0}^* x) + (\alpha_{0,1}^* + \alpha_{1,1}^* x) t + \sum_{k=2}^K (\alpha_{0,k}^* + \alpha_{1,k}^* x) (|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \quad (14)$$

where $\alpha^* = (\alpha_0^* | \alpha_1^*)$ and $\alpha_q^* = (\alpha_{q,0}^*, \dots, \alpha_{q,K}^*)'$ for $q = 0, 1$. Similarly to the model defined for the baseline log-excess hazard, implementation involves a series of transformations to the splines parameters α^* to rewrite the time-dependent effect in terms of these transformed components¹⁵.

Prior distributions

For the Bayesian estimation we choose the following prior distributions for the model parameters:

- For the baseline log-excess hazard as specified in equation (11):

$$\begin{aligned} \alpha_0 &\sim N(0, 10^4), \alpha_1 \sim N(0, 10^4) \\ \alpha_k | \sigma_\alpha &\stackrel{iid}{\sim} N(0, \sigma_\alpha^2), \text{ for } k=2, \dots, K \text{ and } \sigma_\alpha \sim U(0.01, 100) \end{aligned} \quad (15)$$

- For the parameters of a non-linear effect in equation (13):

$$\begin{aligned} \beta_0 &\sim N(0, 10^4) \\ \beta_k | \sigma_\beta &\stackrel{iid}{\sim} N(0, \sigma_\beta^2), \text{ for } k=2, \dots, K \text{ and } \sigma_\beta \sim U(0.01, 100) \end{aligned} \quad (16)$$

- And, for the parameters of a time-dependent effect as in equation (14):

$$\begin{aligned} \alpha_{q,0} &\sim N(0, 10^4), \alpha_{q,1} \sim N(0, 10^4) \text{ for } q=0, 1 \\ \alpha_{q,k} | \sigma_{q,\alpha} &\stackrel{iid}{\sim} N(0, \sigma_{q,\alpha}^2) \text{ for } k=2, \dots, K \text{ and } \sigma_{q,\alpha} \sim U(0.01, 100) \text{ for } q=0, 1 \end{aligned} \quad (17)$$

Measures of interest: excess hazard and net survival

In addition to deriving excess hazard functions and excess hazard ratios for different sets of characteristics of the population, another main quantity of interest that can be derived from an excess hazard model is net survival. Net survival measures the survival in a cohort of cancer patients while considering that the patients can only die from the cancer of interest. A common assumption made when estimating net survival is that the censoring process is non-informative, i.e. the censoring process is independent from the one that generates the events. The process becomes informative when a variable influences both mortality hazards (the cancer-specific and the other-causes mortality hazard), leading to biased estimates of net survival. For example, older patients are more likely to be censored, because of other causes of death, than younger patients, making the censoring process informative. It has been shown that in order to obtain an unbiased estimate of net survival from an excess hazard regression model, the variables that define the population-life tables (from which the other-cause mortality is obtained), and that can influence the censoring process, should be included in the excess hazard model formulation, even if they are not the main focus of the analysis¹⁶. In population-based cancer research, one of the main variables known to influence the censoring process is age at diagnosis, and thus it is advisable to include it in all the log-excess hazard model formulations. It is also advisable to include other variables in the model formulation, such as socio-economic status or region of residence, if life-tables stratified by these variables are available for the population being studied.

Bayesian estimation

After setting up a model for the log-excess hazard, possibly using a combination of several covariate effects modelled using LRTP splines, as specified in the previous sections, the resulting log-likelihood function retains tractability, and thus numerical integration techniques are not needed during the estimation process. Markov Chain Monte Carlo (MCMC) techniques are used to sample from the posterior distributions of all the model parameters. After model convergence has been assessed by inspecting trace and density plots for each parameter, the saved parameter samples are used in a post-estimation procedure to derive posterior distributions of the quantities of interest that can be obtained from excess hazard models. We present post-estimation set-ups to derive posterior distributions of: i) excess hazards, ii) excess hazard ratios for different combinations of population characteristics and iii) net survival for the whole population and for sub-groups of the population.

i) Deriving posterior distributions of excess hazards

Figure 1 shows schematically the post-estimation set-up to derive posterior distributions of excess hazards for different combinations of population characteristics.

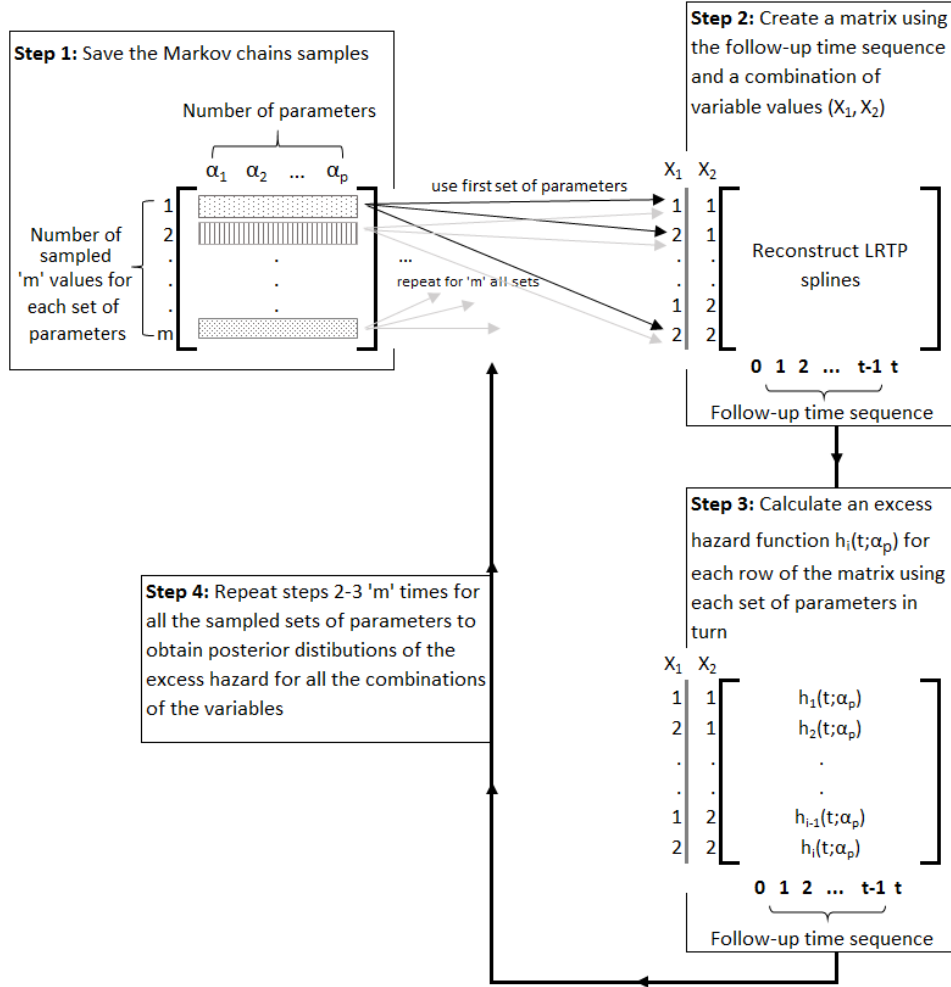


Figure 1. Step-by-step set-up to derive the posterior distributions of excess hazards.

The procedure can be summarised in the following steps:

Step 1 Create a matrix that saves for each parameter (say generically $\alpha_i, i = 1, \dots, p$) the number of sampled Markov chain values (say 'm') from their corresponding posterior distributions.

Step 2 Create a follow-up time sequence within the observed range of follow-up time (0,...,max(t)), and create a matrix containing this time sequence and a combination of values from the variables

entered in the model, chosen within the observed range of values for each variable (e.g. variables X1 and X2 in Figure 1); Re-construct the LRTP splines as defined in the model using the values in the matrix, both for the baseline log-excess hazard and all the covariate effects modelled with LRTP splines.

- Step 3** Use the first set of the ' m ' sampled parameters to estimate an excess hazard function for each combination of variables in the matrix using the follow-up time sequence.
- Step 4** Repeat steps 2-3 ' m ' times for all the sets of sampled parameters (in turn) to obtain posterior distributions of the excess hazard functions for all the combinations of variable characteristics.
- Step 5** Summarise the posterior distributions of the excess hazards using the posterior means, 95% credible intervals and other relevant quantiles.

ii) Deriving posterior distributions for excess hazard ratios

Figure 2 shows the post-estimation set-up for deriving posterior distributions of excess hazards ratios.

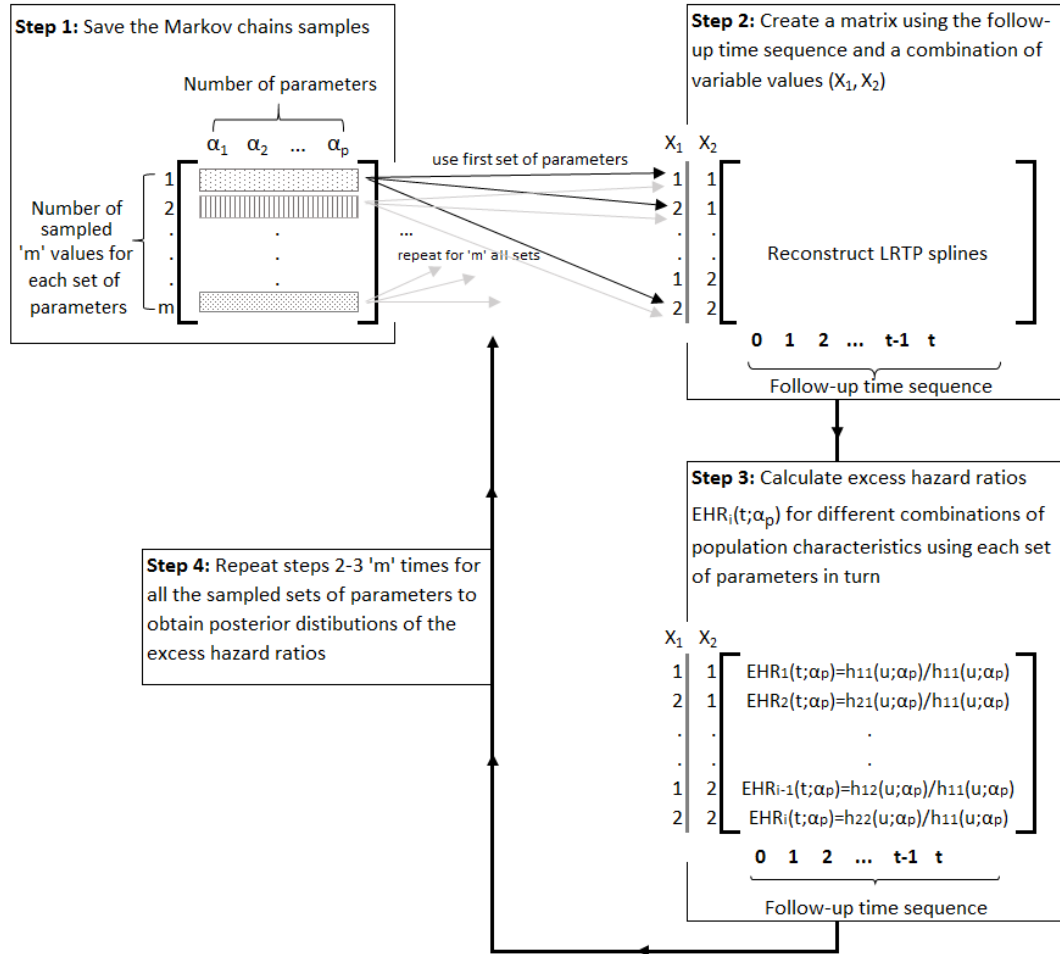


Figure 2. Step-by-step set-up to derive the posterior distributions of excess hazard ratios.

Similarly to the procedure defined in Figure 1 for the excess hazards, the procedure to derive posterior distributions of excess hazard ratios can be summarised as:

Step 1 Same as **Step 1** from the set-up in Figure 1.

Step 2 Same as **Step 2** from the set-up in Figure 1.

Step 3 Use the first set of the 'm' sampled parameters to estimate excess hazard ratios for different combinations of variables in the matrix using the follow-up time sequence.

Step 4 Repeat steps 2-3 'm' times for all the sets of sampled parameters (in turn) to obtain posterior distributions of the excess hazard ratios for all the combinations of variables.

Step 5 Summarise the posterior distributions of the excess hazards ratios using the posterior means, 95% credible intervals and other relevant quantiles.

iii) Deriving posterior distributions of net survival

Figure 3 shows the post-estimation set-up for deriving posterior distributions of net survival for the whole population.

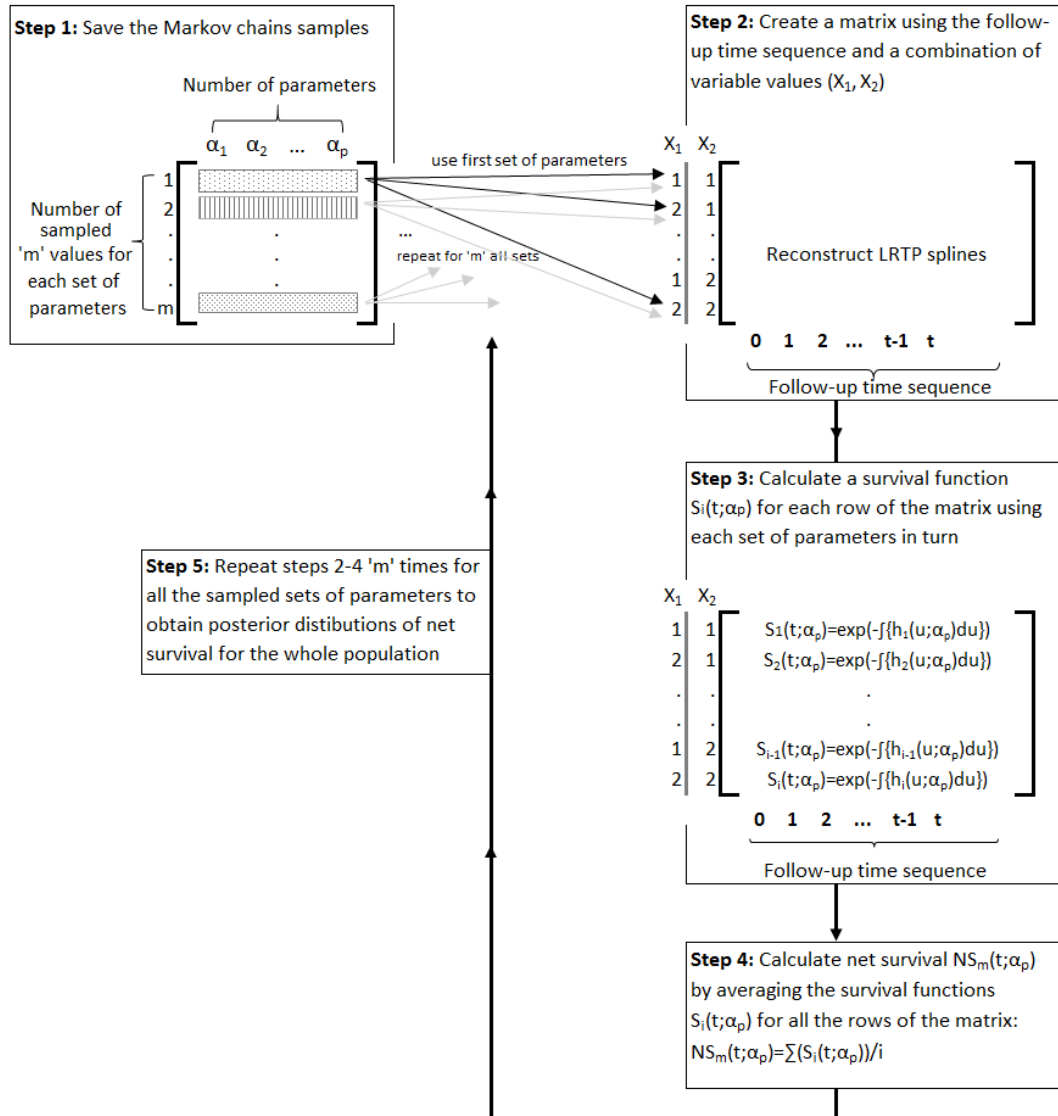


Figure 3. Step-by-step set-up to derive the posterior distributions of net survival for the whole population.

The procedure to derive posterior distributions of net survival can be summarised as:

Step 1 Same as **Step 1** from the set-up in Figure 1.

Step 2 Same as **Step 2** from the set-up in Figure 1.

- Step 3** Use the first set of the ‘ m ’ sampled parameters to estimate a survival function for each entry of the matrix using the follow-up time sequence.
- Step 4** Calculate net survival for the whole population by averaging the survival functions for all the rows of the matrix derived in step 3.
- Step 5** Repeat steps 2-3-4 ‘ m ’ times for all the sets of sampled parameters (in turn) to obtain posterior distributions of net survival for the whole population.
- Step 6** Summarise the posterior distributions of net survival using the posterior means, 95% credible intervals and other relevant quantiles.

The implementation above will provide posterior distributions of net survival for the whole population. We can also derive posterior distributions of net survival by sub-groups of the population, continuing from **step 3** in Figure 3 and averaging the survival functions within each sub-group of the population (e.g. by sub-groups of the variable X_2 as shown by **step 4** in Figure 4).

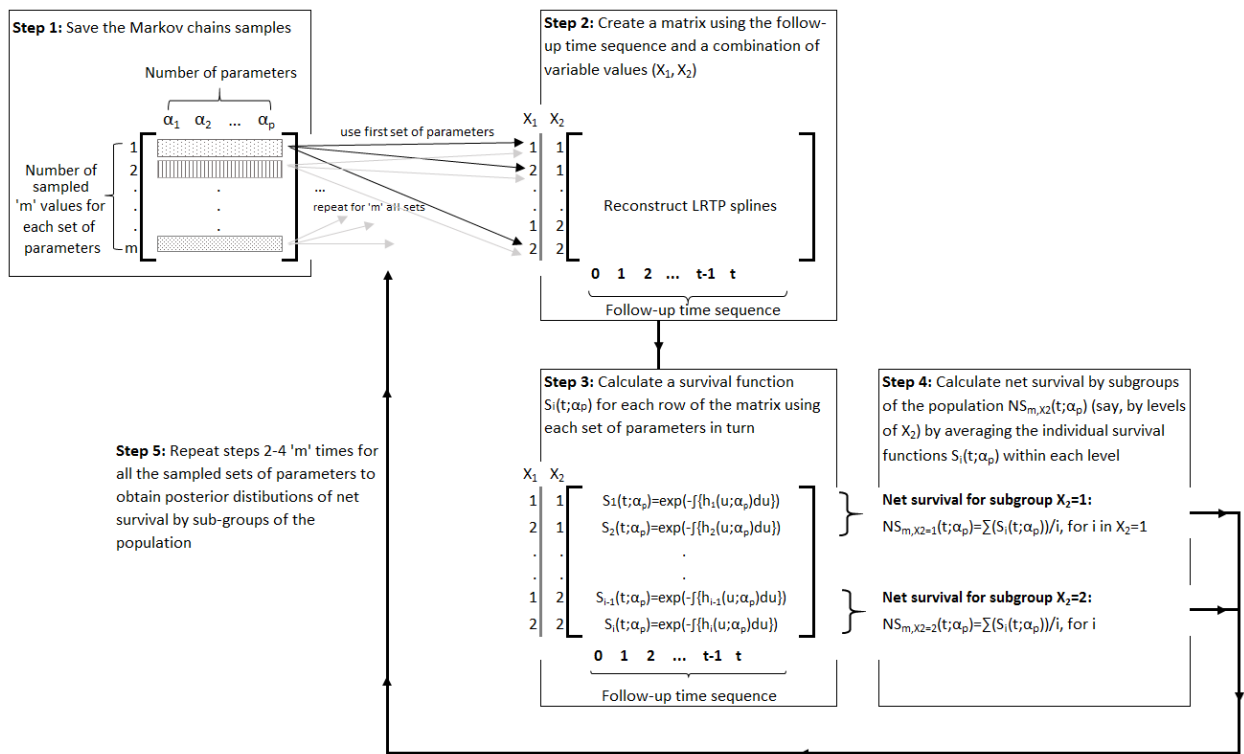


Figure 4. Set-up to derive the posterior distributions of net survival by sub-groups of the population.

Illustration using population-based cancer data

We illustrate the use of the proposed model using data obtained from the National Cancer Registry at the Office for National Statistics (ONS) for all adult men (aged 15-99 years) diagnosed with a first, primary, invasive malignancy of the colon during 2009 in London, England. All patients were followed-up to update their vital status up to six years after diagnosis, until the 31 December 2015. The data variables available for this analysis were: full dates of diagnosis, last follow-up and death, vital-status indicator (dead or censored as alive at the end of follow-up), age at diagnosis (recorded as a continuous variable) and deprivation categories (1-least deprived to 5-most deprived) defined according to the quintiles of the distribution of the Income Domain scores of the 2011 England Indices of Multiple Deprivation¹⁷. Background mortality rates were obtained for each cancer patient from population life tables for England defined for each calendar year in 2009-2015, and stratified by single year of age, sex, deprivation category and region of residence.

Descriptive statistics of the data were performed using the *RStudio* software (version 1.0.153)¹⁸. Bayesian inferences were also performed in *RStudio* using the JAGS MCMC¹⁹ program accessed via the *R* package ‘*R2JAGS*’. R code exemplifying the implementation of the model presented in this illustration is available on the webpage of the *Cancer Survival Group*: <https://csg.lshtm.ac.uk/tools-analysis/>.

The data comprised 1,140 patients. Death was observed for 628 patients (55.1%) over the maximum follow-up period of 5.99 years. Survival time was measured from the date of diagnosis until the date of death or the date of last follow-up. The overall median follow-up time was 3.7 years with standard deviation SD=2.29 years. For patients that died, the median survival time was 0.84 years and for censored patients the median survival time was 5.4 years. The mean age at diagnosis was 70.6 years (SD=13.24 years), and the 25%, 50% and 75% quintiles of the age distribution were 63.2, 72.4 and 80.6 years, respectively. Within deprivation categories, patients were distributed as: 174 (15%) patients in the least deprived category, 207 (18%) patients in the 2nd deprivation category, 223 (20%) patients in the 3rd category, 273 (24%) patients in the 4th category, and 263 (23%) patients in the most deprived category.

A model was set-up for the log-excess hazard including age at diagnosis (A) and deprivation quintile (dep) as main effect covariates. Four partitions ($K=4$) of the observed follow-up time (t) were chosen at the 25%, 50% and 75% percentiles of the events (death) times at $\tilde{t}=(0, 0.18, 0.84, 2.26, 6)$ years. The model can be written as:

$$\begin{aligned}
 \log(h_E(t|\alpha; \beta; \gamma)) &= (\alpha_{0,0} + \alpha_{1,0}A) + (\alpha_{0,1} + \alpha_{1,1}A)t \\
 &+ \sum_{k=2}^K (\alpha_{0,k} + \alpha_{1,k}A)(|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \quad \text{[part 1]} \\
 &+ \beta_1^*(A - \bar{A}) + \sum_{j=2}^J \beta_j^*(|A - \tilde{A}_{j-1}|^3 - |\bar{A} - \tilde{A}_{j-1}|^3) \quad \text{[part 2]} \\
 &+ \gamma * dep \quad \text{[part 3]}
 \end{aligned} \tag{18}$$

where, [part 1] formulates the LRTP spline modelling the baseline log-excess hazard, incorporating the time-dependent effect of age at diagnosis using the same follow-up time partition, with parameters $\alpha = (\alpha_0|\alpha_1)$ and $\alpha_q = (\alpha_{q,0}, \dots, \alpha_{q,K})$ for $q=0,1$. [part 2] represents the LRTP spline modelling the non-linear (smooth) effect of age at diagnosis using 3 partitions ($J=3$) of the observed age range at $\tilde{A}=(16,$

44, 72, 99) years, with parameters β_j , $j = 1, \dots, J$. \bar{A} represents the mean age at diagnosis. For ease of interpretation, age at diagnosis was centered at age 70. [part 3] formulates the linear and proportional effect of deprivation, with parameter γ . This model has 10 parameters associated with the baseline log-excess hazard formulation, including the time-dependent effect of age at diagnosis, and 4 parameters for the regression parameters (3 for the smooth effect of age at diagnosis and 1 for the effect of deprivation). Prior distributions were specified for these parameters using the priors defined in the Methods section, including 3 hyperpriors for the variance parameters of these priors, adding up to a total of 17 model parameters.

The model was fitted setting up 2 MCMC chains, each with 50,000 iterations, a burn-in period of 5,000 and a thinning of 3 to eliminate any existing autocorrelation among samples within the chains. This resulted in a total of 30,000 sampled values from the posterior distributions of each of the 17 parameters. An examination of the trace and density plots of each parameter's posterior distribution did not indicate any convergence issues for these samples. The 30,000 sampled values from the posterior distributions of each parameter, were saved and then used to implement the post-estimation procedure described in Fig. 1 in order to derive posterior distributions for the excess hazard, excess hazard ratios and net survival. Three 'prediction' sequences were created for follow-up time (monthly time points up to five years of follow-up), age at diagnosis (individual integer ages within the observed age range 16-99 years) and deprivation category (1-5). A multi-dimensional matrix was then created to save the results of the posterior distributions for each of the quantities derived, containing the combination of values of all the 'prediction' sequences, and the number of sampled parameter values (30,000). Before the post-estimation procedure was implemented, the splines modelling the baseline log-excess hazard and the smooth effect of age at diagnosis, were reconstructed using the follow-up time and age 'prediction' sequences, maintaining the same spline specification as in the model.

The estimated posterior distributions were summarised by their respective means and other quantiles of interest, such as the 95% credible intervals. For the purpose of this illustration, we present the results in plots (Figures 5-7) showing the mean of each of the posterior distributions.

Interpretation summary: For this cohort of men diagnosed with colon cancer in 2009 in London, England, the estimated mean posterior distributions suggest that: 1) the excess hazard peaks substantially high, up to the first year after diagnosis, for men over 80 years when compared to patients aged 70 years. Whilst for men aged 50 and 60 years their excess hazard is substantially lower, up to the first year after diagnosis, when compared to men aged 70 years. 2); the excess hazard increases gradually for each unit increase in the deprivation category; 3) The mean posterior net survival for the whole cohort shows a moderate decay of the survival curve, reaching approximately 0.6 (60%) at 5 years after diagnosis.

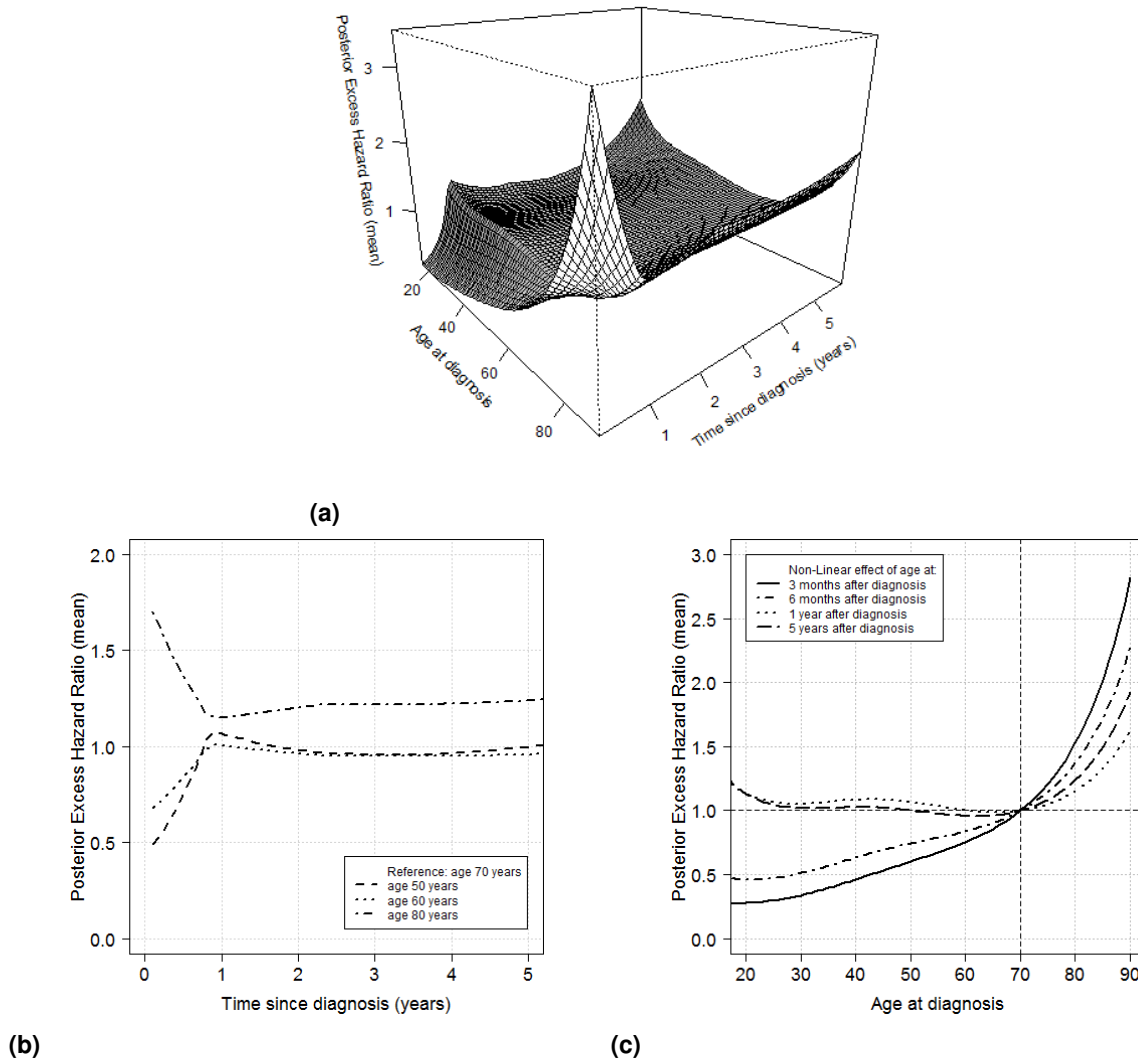


Figure 5. Mean posterior distribution of the excess hazard ratios, showing: (a) a 3-Dimensional representation by age and follow-up time; (b) a slice of the 3-D plot in Fig. 3a) over follow-up time for three ages of diagnosis (50, 60 and 80 years, with 70 years the reference group); (c) a slice of the 3-D plot in Fig. 3a) over age of diagnosis for four follow-up times (3 months, 6 months, 1 year and 5 years after diagnosis).

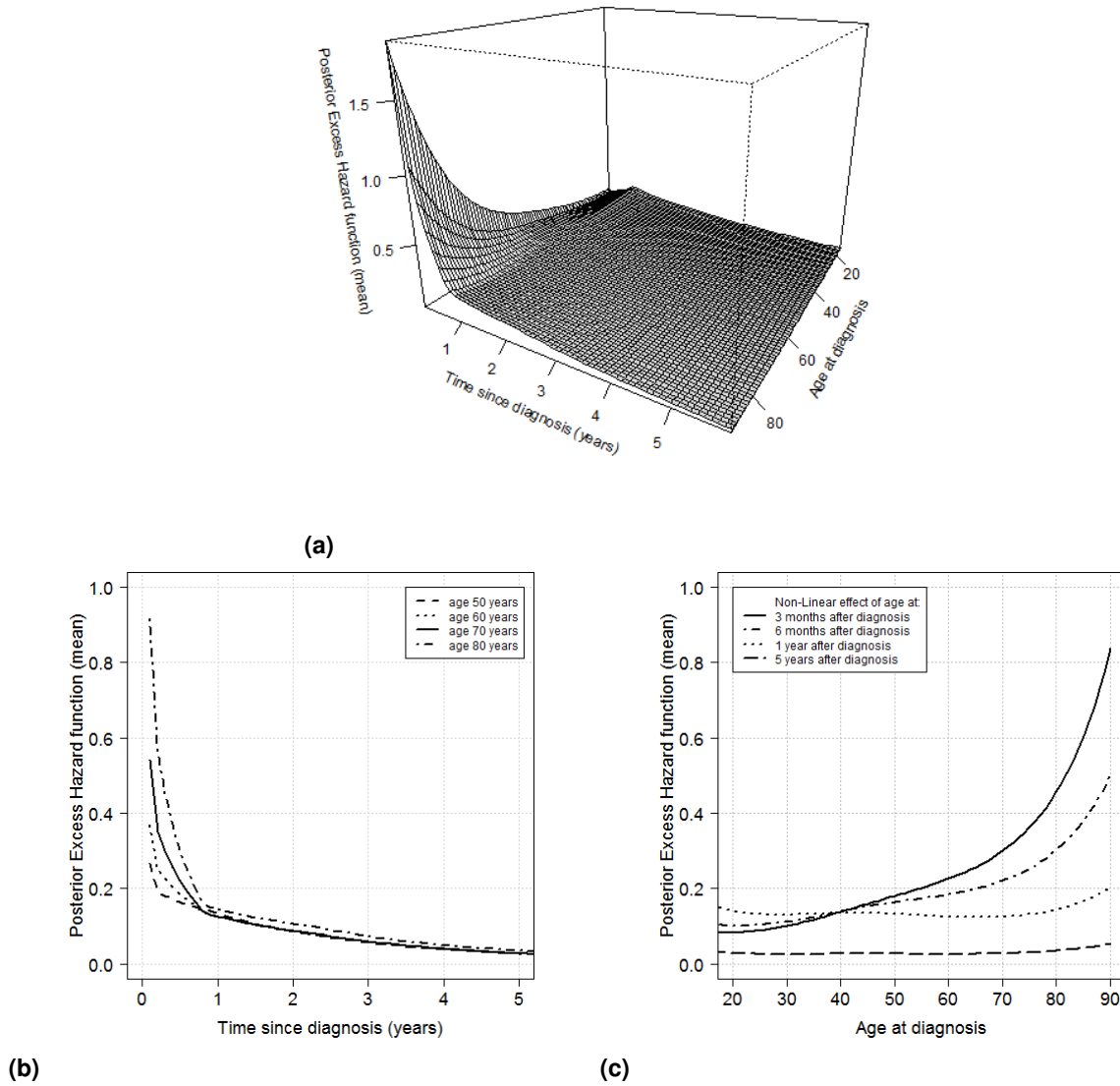


Figure 6. Mean posterior distribution of the excess hazard functions for deprivation category 1 (least deprived patients), showing: (a) a 3-Dimensional representation by age and follow-up time; (b) a slice of the 3-D plot in Fig. 4a) over follow-up time for four age groups (50, 60, 70 and 80 years); and (c) a slice of the 3-D plot in Fig. 4a) over age at diagnosis for four follow-up times (3 months, 6 months, 1 year and 5 years after diagnosis).

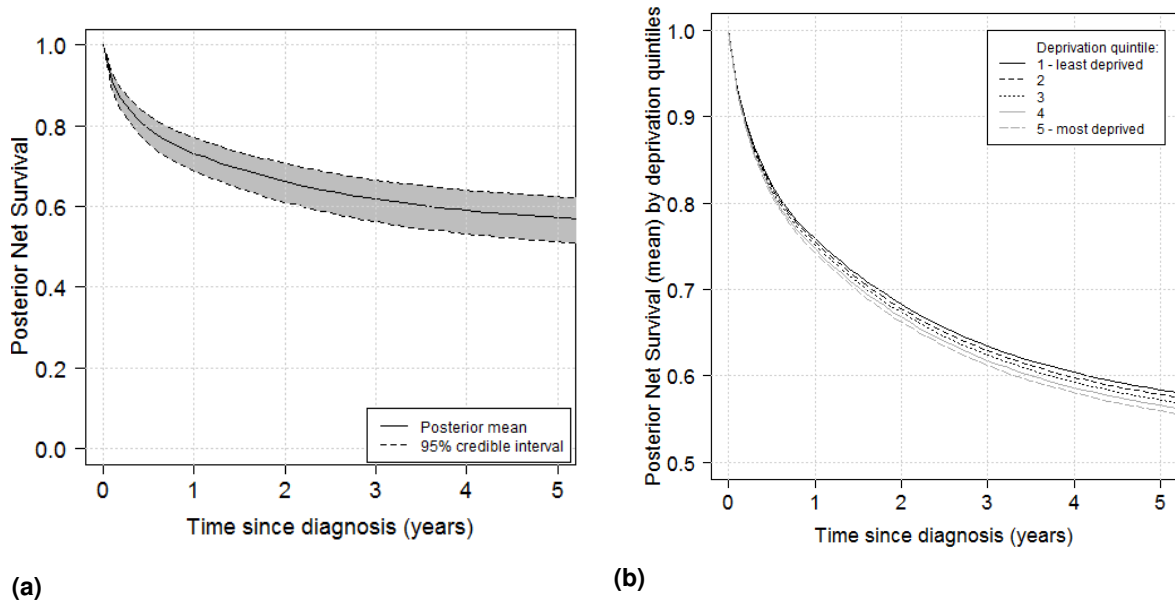


Figure 7. Posterior distribution of net survival, showing: (a) the mean posterior and the 95% credible interval for the whole cohort; and (b) the mean posterior by deprivation category.

Discussion

In this article we introduce a flexible Bayesian regression model for the log-excess hazard, that can be used to investigate inequalities in cancer survival using a range of covariate effects and accommodate different data structures.

Bayesian excess hazard models are very few and none meet our list of criteria set in the Introduction section. Fairly et al.⁸ proposed a model to examine spatial variation in prostate cancer survival using Bayesian relative survival smoothing within a Generalised Linear Model formulation. The number of events was assumed to follow a Poisson distribution, and two random effects were included in the model: a spatially structured random effect for local smoothing and an unstructured random effect global smoothing; Hennerfeind⁹ proposed a Bayesian geoadditive relative survival model using penalized P-splines to model the log-baseline effect as well as the nonlinear and time-varying effects of covariates. Spatial and normal random effects were also included in the model formulation; Cramb et al.¹⁰ introduced a Bayesian flexible parametric model which extends a frequentist flexible parametric model on the log cumulative excess hazard scale using restricted cubic splines²⁰ by adding spatially structured random effects.

We choose here to use Low-Rank Thin Plate splines (LRTP splines) to model the various components of the excess hazard model because they offer a reasonable compromise between model flexibility and

likelihood tractability, with a fast MCMC convergence^{13,14}. Current inference practice for existing log-excess hazard models are mainly done within the frequentist framework, by maximisation of the log-likelihood, and numerical integration techniques are often needed to solve the integral defining the cumulative hazard when flexible functions, such as restricted cubic splines or B-splines, are used to model the different model components. Incorporating higher-dimensional splines would then require to solve numerically extremely complex likelihood functions. Other existing excess hazard models, that are defined on the log cumulative excess hazard scale, have the advantage of avoiding the use of such numerical integration, because of the resulting tractable cumulative excess hazard, but the interpretation of multiple time-dependent effects can be difficult at times when the excess hazard ratio for one variable depends on the levels of the other variables, even without having defined interaction terms in the model²¹.

An additional advantage of using LRTP splines, initially proposed by Murray et al.¹⁵ to model overall hazard, is that their construct is not sensitive to the choice of ‘knot’ location, as is the case with other splines structures, such as restricted cubic splines or B-splines. Murray et al. advise on the selection of a large number of equally spaced partitions of the follow-up time, so that the resulting model can adequately capture the curvature of the hazard function.

In the analysis of the colon cancer data, we selected several partitions of the follow-up time (between 2 and 20), using a mixture of equally spaced and pre-defined intervals. Models were compared using the Deviance Information Criterion (DIC)²⁴, and the model presented in the results section (using 4 partitions of the follow-up time) corresponded to the model with the smallest DIC. We found that less partitions (four in our analysis) did adequately capture the shape of the excess hazard function for the cancer analysed, and that partitioning the event times at the percentiles captured well the largest shift in the decay of the function in the first year after diagnosis. The shapes of the baseline excess hazard function and of the age-related function defined in our final model were also very similar to those estimated by the frequentist flexible excess hazard model using higher-dimensional splines²⁵. We also observed that using less partitions substantially decreased computation time when fitting these models using MCMC sampling (for example, time was reduced by a quarter when using 4 instead of 20 partitions), as there are less parameters to be sampled. We note that fitting these models can be computationally very expensive, varying from a few hours to a few days, depending on model complexity, the number of MCMC iterations and the size of the matrices generated. Computation time can be reduced by the use of parallel computing, the use of computers with Graphics Processing Units (GPU), or by exploring new advances in accelerated computing such as GPU-accelerated packages²⁶.

Eliciting informative priors for the model parameters was not within the aim of this study, and we opted to choose vague priors for all the model parameters. In such a scenario, the mean posterior distributions for the parameters and quantities of interest, would be closer to the Maximum Likelihood estimates obtained using a similar model set-up.

A novel component that this article offers, is the implementation of a post-estimation procedure (as described in Fig. 1), to derive posterior distributions for the excess hazard ratios, excess hazard functions and net survival, based on the saved MCMC samples for each parameter. This procedure, as described, derives posterior distributions using a predefined matrix that contains a combination of values of the covariates within the observed range in the data, and it does not use the data for the whole cohort. The estimation of excess hazards and excess hazard ratios is usually made for different sets of characteristics of the cohort, and thus it is easier to construct a matrix to derive these posterior distributions. For net survival, the estimation is made by averaging the individual survival curves, which can be done using one

of two options: 1) use the whole cohort of observed data, estimate a survival curve for each observation (following the same procedure as outlined in Fig. 1), and then average over the whole cohort to obtain an estimate of net survival, or 2) use the matrix, estimate a survival curve for each combination of values of covariates, and then average over these curves to obtain an estimate of net survival. The main advantage of using a matrix over the observed cohort is the reduced computation time, especially when large cohorts are analysed. In addition, when using a fixed covariate structure within the matrix, the results will be internally standardised for those covariates, and thus when comparing net survival by sub-groups of the cohort, this has the advantage that comparability will already be taken into account. For example, if we consider two variables, age at diagnosis and deprivation, and estimate net survival by deprivation category, averaging the individual survival curves within each deprivation group using the whole cohort, if the age distribution within deprivation category is very different, the results will not be comparable. But if we use a matrix with a fixed age structure for all levels of deprivation, the estimated net survival curves will be comparable between deprivation categories.

One of the criteria we set up a priori for the implementation of the proposed model, was that it could be easily extended to include one or more random effects to accommodate clustered data, and incorporate hierarchical data structures. The Bayesian framework lends itself very nicely to specify models with such characteristics. We have extended the model specified in equation (18) to add two random effects: one clustering patients by area of residence, and another clustering patients by treatment center. Although the model implementation was a straight-forward step from the previous model implementation (without random effects), we found some convergency problems when using the open-source MCMC sampler, and a substantial increase in computation time, depending on the size and number of clusters used (results not shown). We propose as further extension to this work, to develop a dedicated MCMC sampler that improves sampling from the parameters' posterior distributions when using these more complex model structures.

In summary, we have shown how a flexible Bayesian model for the log-excess hazard can be used for population-based research, to investigate socio-economic inequalities in cancer survival using a range of covariate effects modelled using LRTP splines. In our experience, we found that using LRTP splines provides a good compromise between the achieved model flexibility and the retained tractability that reduces computational intensity. Although constructing these splines involves many matrix calculations in order to compute the necessary transformations to implement the splines, the user-friendly and modifiable code that has been made available¹⁵ makes the implementation uncomplicated. In particular, we think that the new post-estimation process we propose to derive posterior distributions for net survival and excess hazards will be a very useful tool for cancer researchers in the production of cancer survival statistics with relevance to health policy.

Acknowledgements

The authors wish to acknowledge Dr. Francisco Rubio for the very helpful discussions and encouragements.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first and last authors wish to thank Cancer Research UK for the funding provided for this research, grant numbers C1336/A11700 and C7923/A18348. The second author wishes to thank the UK Medical Research Council for the funding provided for this research, grant numbers MC_UU_12023/21 and MC_UU/12023/29.

References

1. Estève J, Benhamou E, Croasdale M et al. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*. 1990; **9**: 529-538.
2. Remontet L, Bossard N, Belot A et al. An overall strategy based on regression models to estimate relative survival and models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*. 2007; **26**: 2214-2228.
3. Lambert PC and Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal*. 2010; **9**: 265-290.
4. Perme MP, Stare J and Estève J. On estimation in relative survival. *Biometrics*. 2012; **68**: 113-120.
5. Perme MP, Estève J and Rachet B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer*. 2016; **16**.
6. Crowther MJ and Lambert PC. A general framework for parametric survival analysis. *Statistics in Medicine*. 2014; **33**: 5280-5297.
7. Giorgi R, Sahel A, Daures JP et al. A Metropolis within Gibbs sampling in relative survival. *Far East Journal of Theoretical Statistics*. 2005; **16**: 269-284.
8. Fairley L, Forman D, West R et al. Spatial variation in prostate cancer survival in the Northern and Yorkshire region of England using Bayesian relative survival smoothing. *British Journal of Cancer*. 2008; **99**: 1786-1793.
9. Hennerfeind A, Held L and Sauleau EA. A Bayesian analysis of relative cancer survival with geoadditive models. *Statistical Modelling*. 2008; **8**: 117-139.
10. Cramb SM, Mengersen KL, Lambert PC et al. A flexible parametric approach to examining spatial variation in relative survival. *Statistics in Medicine*. 2016; **35**: 5448-5463.
11. Collett D. Modelling Survival Data in Medical Research. Chapman & Hall, 2nd ed. 2003.
12. Rachet B, Maringe C, Woods LM et al. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health*. 2015; **15**: 1240.
13. Ruppert D, Wand MP and Carroll RJ. Semiparametric Regression. *Cambridge Series in Statistical and Probabilistic Mathematics*. 2003.
14. Crainiceanu CM, Ruppert D and Wand MP. Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*. 2005; **14**: 1-24.
15. Murray TA, Hobbs BP, Sargent DJ et al. Flexible Bayesian Survival Modeling with Semiparametric Time-Dependent and Shape-Restricted Covariate Effects. *Bayesian Analysis*. 2016; **11**(2): 381-402.
16. Danieli C, Remontet L, Bossard N et al. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*. 2012; **31**(8): 775-86.
17. English indices of deprivation. *Ministry of Housing, Communities and Local Government*. 2011. URL <https://www.gov.uk/government/collections/english-indices-of-deprivation>.
18. RStudio Team. RStudio: Integrated Development for R. 2015. URL <http://www.rstudio.com/>.
19. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.

20. Nelson CP, Lambert PC, Squire IB et al. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*. 2007; **26**: 5486-5498.
21. Royston P and Lambert PC. Flexible Parametric Survival Analysis using Stata: Beyond the Cox Model. *Stata Press*. First edition, 2011.
22. Klein JP and Moeschberger ML. Survival analysis: techniques for censored and truncated data. *Springer*, 2003.
23. Brenner H, Gefeller O and Hakulinen T. Period analysis for up-to-date cancer survival data. Theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer*, 2004; **40**: 326-335.
24. Spiegelhalter DJ, Best N and Carlin BP and van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 2002; **64**(4): 583-639.
25. Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, Launoy G, Belot A and CENSUR Working Survival Group. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 2016; **35**: 3066-84.
26. Draper D and Terenin A. Comment: A brief survey of the current state of play for Bayesian computation in data science at big-data scale. *Brazilian Journal of Probability and Statistic*. 2017; **31**(4): 686–69.

6.5 Research publication 4

Title: 'Variation in colon cancer survival for patients living and receiving care in London, 2006-2013: does where you live matter?'.

Authors: Manuela Quaresma, James Carpenter, Adrian Turculet and Bernard Rachet.

Manuscript prepared for submission to *The Lancet* inserted from next page.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	199304	Title	Ms
First Name(s)	Manuela		
Surname/Family Name	Quaresma		
Thesis Title	Population-based cancer survival at small area level: methodological developments		
Primary Supervisor	Professor Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Lancet Oncology
Please list the paper's authors in the intended authorship order:	Manuela Quaresma, Adrian Turculet, James Carpenter Bernard Rachet
Stage of publication	Not yet submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	MQ developed the application and run the analysis supervised by JC and BR. AT produced the maps in ArcGIS and the windroses graphs. MQ wrote the article. MQ, JC and BR commented on the structure and revised the article.
--	---

SECTION E

Student Signature	
Date	20/11/2019

Supervisor Signature	
Date	20/11/2019

Title: Variation in colon cancer survival for patients living and receiving care in London, 2006-2013: does where you live matter?

Authors: Manuela Quaresma¹, James Carpenter^{1,2}, Adrian Turculet¹, Bernard Rachet¹

Abstract

Marked geographical variations in cancer survival have been consistently described for most common adult cancers in England. Similar patterns have been observed within the capital London, almost mimicking a microcosm of the country's survival patterns. This evidence has suggested that the place of residence might play an important role in the survival of cancer patients. In this study, we analysed data for patients diagnosed with colon cancer, who were living within a London Clinical Commissioning Group (CCG) at the time of their diagnosis and received cancer care in a hospital located within a London CCG. We investigated the patterns of patient pathways between the CCG of residence and the hospital of cancer care, and estimated the variability in survival at both CCG and hospital level. The most frequent pathway patients travelled was to the hospitals located closest to their area of residence. After adjusting for age at diagnosis, socioeconomic status, stage at diagnosis and hospital of care, no variability in survival was observed between CCGs. This result contrasted with a much more pronounced variability between hospitals. This study demonstrates the importance of performing more in-depth investigations into the disparities in cancer survival using population cancer data enriched with other relevant electronic health data sources.

Keywords: colon cancer, net survival, variation, London, CCG, hospital

¹London School of Hygiene & Tropical Medicine, Faculty of Epidemiology & Population Health, London, UK

²London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, UK

Introduction

Population-based cancer survival statistics provide key insights into the overall effectiveness of a healthcare system in managing and treating cancer. Quantifying disparities in cancer survival in particular can directly identify areas of inequity amenable to change. For instance, wide geographical and socioeconomic inequalities in cancer survival have been consistently described, despite the existence of universal access to care within the National Health Service (NHS), founded on the principles of equity and free access to all. A clear and persistent North-South gradient, with lower survival in the North of England, exists for most common adult cancer types, while similar patterns are observed within London, almost mimicking a microcosm of the country's survival patterns. This evidence has suggested that the place of residence might play an important role in the survival of a cancer patient, giving rise to much political debate since the introduction of the first NHS cancer plan and other national initiatives aimed at tackling cancer inequalities. Following the 2012 Health and Social Care Act and the subsequent restructuring of the NHS, two organisations became central role players in the organisation and commissioning of care: NHS England and the (now 211) Clinical Commissioning Groups (CCGs). NHS England became responsible for commissioning the planning and buying of health care services, such as primary care services, and setting the priorities and direction of the NHS. It also allocates 60% of the NHS budget to CCGs across England. CCGs are clinically led statutory NHS bodies, responsible for the planning and commissioning of healthcare services for their local area, including General Practitioner (GP) services, planned hospital, urgent and emergency care. Cancer survival outcomes for CCGs have been published on a regular basis since their creation, including an index of cancer survival for all cancers combined and cancer-specific survival indexes for breast, colorectum and lung cancers. The CCG outcomes continue to support previous evidence of wide variation in survival across England, including large variation between CCGs within London. Understanding the mechanisms underlying such wide disparities, requires addressing multiple research questions to disentangle the different aspects of the multi-layered and multi-factorial 'cancer inequalities puzzle', including the integrated study of patient-, tumour- and health-system characteristics.

In this article we analyse data for patients diagnosed with colon cancer, who were living within the catchment area of a London CCG at the time of their diagnosis and received cancer care in a hospital located within a London CCG. We start by investigating the patterns of patient pathways between the area of residence and the hospital of cancer care. We then investigate the variability in cancer survival at both CCG and hospital level, after adjusting for some patient and tumour characteristics, such as age at diagnosis, socioeconomic status and stage at diagnosis.

Material and Methods

Data

Data on individual cancer records were obtained from the National Cancer Registry at the Office for National Statistics (ONS) for all adults (aged 15-99 years), diagnosed with a first, primary, invasive malignancy of the colon during 2006-2013 in London, England. All patients were followed-up to update their vital status until the 31st December 2014. The data variables available for analysis from this data source were: gender, age at diagnosis, full dates of diagnosis, last follow-up and death, vital status indicator (dead or censored as alive at the end of follow-up), CCG of residence at diagnosis, deprivation category (1-least deprived to 5-most deprived) and stage at diagnosis (1-localised cancer stage to 4-metastatic cancer stage). A CCG of residence was allocated to each patient based on their postcode of residence. Since CCGs only came into existence in 2013, for coherence in the analysis, we applied the CCG boundaries retrospectively to patient records diagnosed prior to 2013 based on historical postcode files. Patients were also allocated to one of five deprivation categories at the time of their diagnosis using the Income Domain from the 2011 England Indices of Multiple Deprivation defined at the Lower Super Output Area level (LSOA). To complement the cancer registry dataset with information on stage at diagnosis and hospital of cancer care, each individual cancer record was linked to two additional sources of data, Hospital Episode Statistics (HES) records and the national bowel cancer clinical audit data (NBOCA) using a data linkage algorithm by Shack et al. (1). After the three data sources were linked, the stage at diagnosis variable was reconstructed using

the algorithm by Benitez Majano et al. (2) that combines available information on tumour (T), nodes (N) and metastases (M). The algorithm prioritises information captured in the clinical audit data and if not available uses cancer registry stage data. Treatment information was also derived from clinical audit data and HES records using an algorithm by Fowler et al. (3) that categorises major surgical treatment received by each patient within a time window of between 30 days prior and 90 days following cancer diagnosis, and categorises other minor forms of treatment (including palliative care and diagnostic procedures if no other treatment was recorded) into a minor treatment category. Based on the previous definition of treatment categories, we allocated to each cancer patient a hospital of cancer care, or of diagnosis if no major surgical treatment was received, using a combination of different variables available in the data containing hospital codes.

Statistical methods and data visualisation

In addition to usual descriptive statistics, various data visualisation techniques were used. Windrose graphs were used to display the distribution of patients' deprivation category and stage at diagnosis by CCG of residence and hospital of cancer care. CCGs and hospitals were arranged in the windroses according to their approximate cardinal directions of location in London for ease of visualisation. Flow maps of London were created to visualise patterns of patient pathways between the CCG of residence and the hospital of cancer care. The maps show the areas of catchment and boundaries for each of the 32 London CCGs, all identified with their names. The 36 London hospitals used in this study are marked on the maps using the exact location based on their latitude and longitude coordinates. The key to the hospital names is given in the map legend using the identifiers ($H1, H2, \dots, H36$). Each pathway is shown on the map using lines connecting the centroid of each CCG (black dot) to each hospital. The pathway line colours distinguish between the frequency of each pathway, coloured from the most frequent up to the 5th most frequent, with the proportion (%) of patients using each pathway indicated on the lines. Only pathways that had more than 5% of patients were drawn and thus the sum of all the pathway frequencies originating from each CCG will not add to 100%. Maps were created using the software ArcGIS 10.5 (5).

In order to investigate the variability in cancer survival at CCG and hospital levels, net survival (survival from the cancer) and excess hazards of death (hazards due to the cancer) were estimated using flexible Bayesian excess hazard models proposed by *Quaresma et al.* (6). Separate models were fitted for men and women, adjusting for age at diagnosis, deprivation category and stage at diagnosis. To accommodate the hierarchical structure of the data (i.e. that patients within a given CCG of residence or hospital of cancer care are likely to share some characteristics), the original model by Quaresma et al. was extended with the inclusion of a pair of random effects for CCG and hospital. To isolate the excess (cancer-related) hazards of death, the hazards of death from other causes were obtained for each cancer patient from English life tables defined for each calendar year in 2006-2014 and stratified by single year of age, sex, deprivation category and region of residence (7,8). Five-year net survival for each CCG and hospital was estimated (based on the mean of their posterior distributions) and their variability across CCGs and hospitals was presented using funnel plots (9). Details on the complete model specification, including a model extension to handle the missing information on stage at diagnosis are given in Appendix A3. Conventional analyses were completed using Stata 15 (4) whereas Bayesian inferences were performed in R software version 3.4.3 using the JAGS MCMC program accessed via the R package 'R2JAGS' (10,11).

Results

Data were available on 16,326 patients diagnosed with colon cancer between 2006-2013 in London, England (see flow chart in Figure 6.1). For 15,309 (94%) patients, a hospital of cancer care was successfully allocated after the treatment capture algorithm was applied to each cancer record. The 1,017 (6%) patients for which a hospital of cancer care or diagnosis could not be allocated were not included in further analyses. For 10,869 (71%) of the eligible 15,309 patients, the hospital allocated corresponded to the hospital where the patient underwent a major surgery for colon cancer. For the remaining 4,440 (29%) patients, the hospital allocated corresponded either to the hospital of diagnosis provision or palliative care, if no major surgical treatment was recorded.

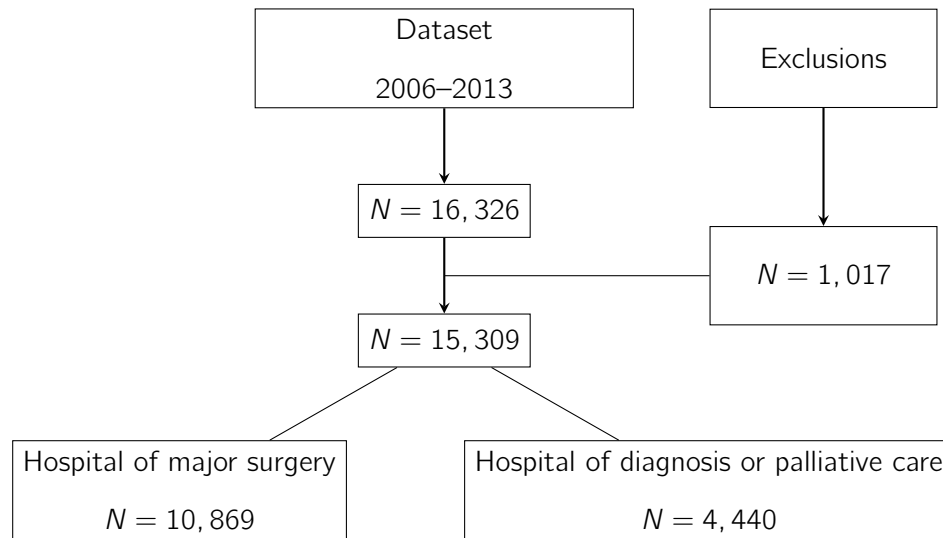


Figure 6.1: Flow chart of data exclusions and hospital assignment after applying the algorithm to allocate the hospital of care or diagnosis.

Individual characteristics of colon cancer patients by CCG and hospital

Tables 6.1 and 6.2 show the distribution of cases and deaths for men and women by CCG of residence and hospital of cancer care, respectively. Of the 15,309 patients included in the analysis, 7,841 (51%) were men and 7,468 (49%) were women. Death was observed for 7,674 (50%) patients over the maximum follow-up period of 8.9 years. Deaths ranged between 40–60% in both men and women for CCG of residence, and between 30–77% in men and 38–75% in women for hospital of care. Survival time was measured from the date of diagnosis until the date of death or the date of last follow-up. For patients that died, the median survival time was 0.72 years and for censored patients the median survival time was 4.1 years. The mean age at diagnosis was 72 years (SD=13.2) for men and 74 years (SD=14.4) for women. For both men and women, the overall distribution of patients within deprivation categories was similar, ranging from 13% of patients in the least deprived group to 27% in the most deprived group. Stage at diagnosis was missing for 23% of the cases. Among the records with observed stage, the overall stage distribution was similar for both men and women, with 13% of patients diagnosed with stage 1 disease, 34% with stage 2, 34% with stage 3 and 19% of patients diagnosed with stage 4. The windrose graphs show that the highest proportion of patients from the most deprived group came from the North East/East London CCGs and hospitals, reaching over 80% of patients in some areas

compared to the South West/South London areas where patients from the least deprived group are more predominant, although in much smaller proportions (Figures 6.2 a) and 6.2 b)). The distribution of stages 1, 2 and 3 (grouped into one category) ranged between 50-73% by CCG of residence and between 37-80% by hospital of care. The distribution of patients with stage 4 was similar by CCG and hospital, ranging between 6-26% (Figures 6.2 c) and 6.2 d)). These patterns were similar both for men and women. Additional interactive windroses charting the distribution of all deprivation categories and all stages at diagnosis by CCG and hospitals, and equivalent graphs for the distribution of women can be accessed via <https://csg.lshtm.ac.uk/survival-variation-CCG-hospital-London/>.

Pathways of colon cancer patients between their CCG of residence and hospital of cancer care

The flow maps in Figures 6.3 and 6.4 display the pathways of patients between the CCG of residence and the hospital of cancer care for men and women, respectively. Overall, the most frequent pathway patients travelled was to the closest hospital located within the catchment area of their CCG of residence. Similar pathway frequencies were observed for both men and women. Three main patterns can be distinguished: a) For one third of CCGs, namely Bromley, City and Hackney, Croydon, Greenwich, Havering, Hillingdon, Hounslow, Kingston, Newham, Waltham Forest and Tower Hamlets, more than 70% of patients travelled to one main hospital closest to their area of residence, and with lower frequency to other hospitals. In particular, for patients living in Waltham Forest and Tower Hamlets (and Kingston for women) more than 90% travelled to only one hospital. b) The second pattern identified CCGs in which patients travelled with similar frequency to two main hospitals close to their areas of residence, namely Barking and Dagenham, Bexley, Camden, Islington. c) For the remaining 17 CCGs, patients travelled more frequently up to three or four hospitals, travelling further to hospitals outside of their CCG of residence. Overall, the patterns displayed in the flow maps clearly define areas in London where patients' travels are more self-contained to hospitals located in their neighbouring areas, such as for example, in the North East, East and South East of London. In contrast, patients living in the North

and South West of London tend to access more hospitals outside their area of residence, most of them located in central London.

Variations in five-year colon cancer survival

Posterior distributions of five-year net survival were derived for each CCG of residence and hospital of cancer care from the multivariable excess hazard model, which included in addition to CCG and hospital, age at diagnosis, deprivation and stage (full model). Complete model specification and Bayesian inference details are presented in Appendixes A1-A4. From these posterior distributions, funnel plots were created by CCG of residence and hospital of care (Figures 6.5 and 6.6 for men and women, respectively). Each funnel plot charts 5-year net survival (posterior mean) against their corresponding precisions. Superimposed on the funnel plots are the 95% and 99.8% control limits. The target values (horizontal lines) were taken as the mean net survival for London. Plots were presented stratified by stage at diagnosis because the level of survival is very differential between early stages (stages 1, 2 and 3) and late stage (stage 4). No variability was observed between CCGs for both men and women (Figures 6.5 a), 6.5 c), 6.6 a) and 6.6 c)), with all estimates almost exactly at the same level as the target line. However, large variability was observed between hospitals, although most of the estimates were contained within the 99.8% control limits in the funnel plots. For stages at diagnosis 1, 2 and 3, hospital-specific five-year net survival ranged between 61-77% for men (with target 69%) (Figure 6.5 b)) and between 67-76% for women (with target 72%) (Figure 6.6 b)). For stage at diagnosis 4, the survival estimates ranged between 10-28% for men (with target 18%) (Figure 6.5 d)) and between 19-32% for women (with target 26%) (Figure 6.6 d)).

For comparison of results with the full model, three additional excess hazard models were fitted by adding covariates successively: Model 1, including age and CCG; Model 2, including age, CCG and deprivation; Model 3, including age, CCG, deprivation and stage. Based on each of these models, funnel plots were created by CCG of residence to visualise if any survival variability by CCGs was observed before the fully adjusted model. For both men and women, five-year net survival varied moderately between CCGs, even after adjusting for age at diagnosis, deprivation and stage at diagnosis (Figures 6.7 a) b) c), Figures 6.8

a) b) c), Figures 6.9 a) b) c) and Figures 6.10 a) b) c)). Such disparities disappeared once adjusted for hospital of cancer care as shown by the funnel plots in Figures 6.7 d), 6.8 d), 6.9 d) and 6.10 d).

CCG of residence	Men		Women	
	Cases (N)	Deaths (%)	Cases (N)	Deaths (%)
C1: Barking and Dagenham	194	58.8	200	53.5
C2: Barnet	397	48.9	326	48.8
C3: Bexley	308	52.9	270	51.5
C4: Brent	263	44.9	240	49.2
C5: Bromley	414	54.1	428	53.3
C6: Camden	172	51.7	177	46.9
C7: Central London	156	54.5	109	41.3
C8: City and Hackney	187	49.7	176	51.1
C9: Croydon	401	46.9	369	50.7
C10: Ealing	308	50.0	303	47.2
C11: Enfield	308	51.3	323	49.8
C12: Greenwich	246	52.0	223	47.1
C13: Hammersmith and Fulham	154	48.7	160	46.2
C14: Haringey	212	50.9	204	52.9
C15: Harrow	234	42.7	221	44.8
C16: Havering	356	55.3	360	53.6
C17: Hillingdon	310	54.5	298	49.7
C18: Hounslow	215	41.9	206	50.0
C19: Islington	183	52.5	168	44.0
C20: Kingston	173	47.9	197	52.3
C21: Lambeth	240	44.2	245	47.3
C22: Lewisham	221	49.3	215	53.0
C23: Merton	218	48.6	217	50.2
C24: Newham	181	53.6	142	48.6
C25: Redbridge	268	50.4	275	50.9
C26: Richmond	252	43.6	225	46.7
C27: Southwark	218	50.9	214	54.2
C28: Sutton	254	45.7	253	49.0
C29: Tower Hamlets	139	59.7	142	54.9
C30: Waltham Forest	206	51.9	202	58.4
C31: Wandsworth	270	50.4	218	51.8
C32: West London	183	53.5	162	40.1
Total	7,841	50.3	7,468	50.0

Table 6.1: Number of cases (N) and proportion of deaths (%) within the follow-up period by CCG of residence for men and women diagnosed with colon cancer in London, 2006-2013.

Hospital of cancer care	Men		Women	
	Cases (N)	Deaths (%)	Cases (N)	Deaths (%)
H1: Barnet hospital	234	47.9	195	50.8
H2: Central Middlesex hospital	48	68.7	48	75.0
H3: Charing Cross hospital	169	49.7	177	43.5
H4: Chase Farm hospital	182	59.3	193	49.2
H5: Chelsea and Westminster hospital	180	51.7	180	42.2
H6: Croydon University hospital	319	50.8	320	53.4
H7: Ealing hospital	181	52.5	153	52.9
H8: Epsom hospital	85	29.4	82	37.8
H9: Guy's hospital	88	37.5	90	42.2
H10: Hammersmith hospital	72	65.3	68	55.9
H11: Hillingdon hospital	245	57.9	247	51.4
H12: Homerton University hospital	166	50.0	153	51.6
H13: King George hospital	241	58.1	233	55.8
H14: King's College hospital	271	48.7	253	50.2
H15: Kingston hospital	354	48.0	355	50.7
H16: Mount Vernon hospital	30	76.7	18	61.1
H17: Newham General hospital	142	57.7	128	50.8
H18: North Middlesex hospital	187	55.6	176	57.4
H19: Northwick Park hospital	236	42.4	209	46.9
H20: Princess Royal University hospital	372	55.1	368	54.9
H21: Queen Elizabeth hospital	339	52.2	277	49.1
H22: Queen Mary's hospital	178	57.9	187	55.6
H23: Queen's hospital	435	55.4	469	54.2
H24: Royal Free hospital	235	51.5	217	49.8
H25: St. George's hospital	377	38.2	320	40.0
H26: St. Helier hospital	234	58.9	244	59.8
H27: St. Mark's hospital	229	37.1	237	38.8
H28: St. Mary's hospital	265	39.2	227	38.3
H29: St. Thomas' hospital	247	55.1	230	53.9
H30: The Royal London hospital	210	52.8	183	51.9
H31: The Royal Marsden hospital	99	40.4	99	45.5
H32: The Whittington hospital	181	56.3	199	47.2
H33: University College hospital	283	40.9	244	39.3
H34: University hospital Lewisham	181	47.5	180	49.4
H35: West Middlesex University hospital	237	44.7	213	54.4
H36: Whipps Cross hospital	309	50.8	296	53.4
Total	7,841	50.3	7,468	50.0

Table 6.2: Number of cases (N) and proportions of deaths (%) within the follow period by hospital of cancer care for men and women diagnosed with colon cancer in London, 2006-2013.

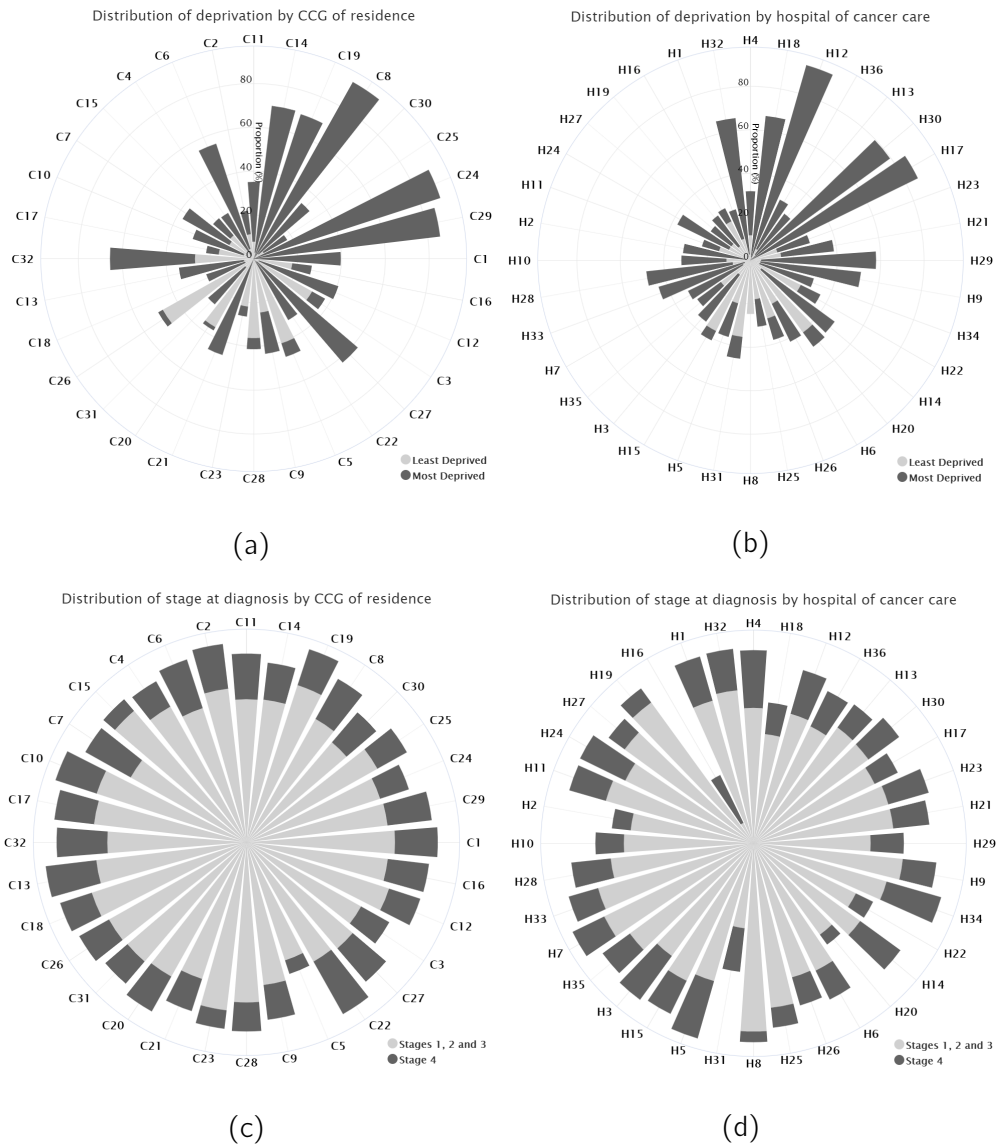


Figure 6.2: Windrose graphs showing the distribution (%) of male patients diagnosed with colon cancer in London, 2006-2013: (a) least deprived versus most deprived category by CCG of residence; (b) least deprived versus most deprived category by hospital of cancer care; (c) stages at diagnosis 1, 2 and 3 versus stage 4 by CCG of residence (d) stages at diagnosis 1, 2 and 3 versus stage 4 by hospital of cancer care.

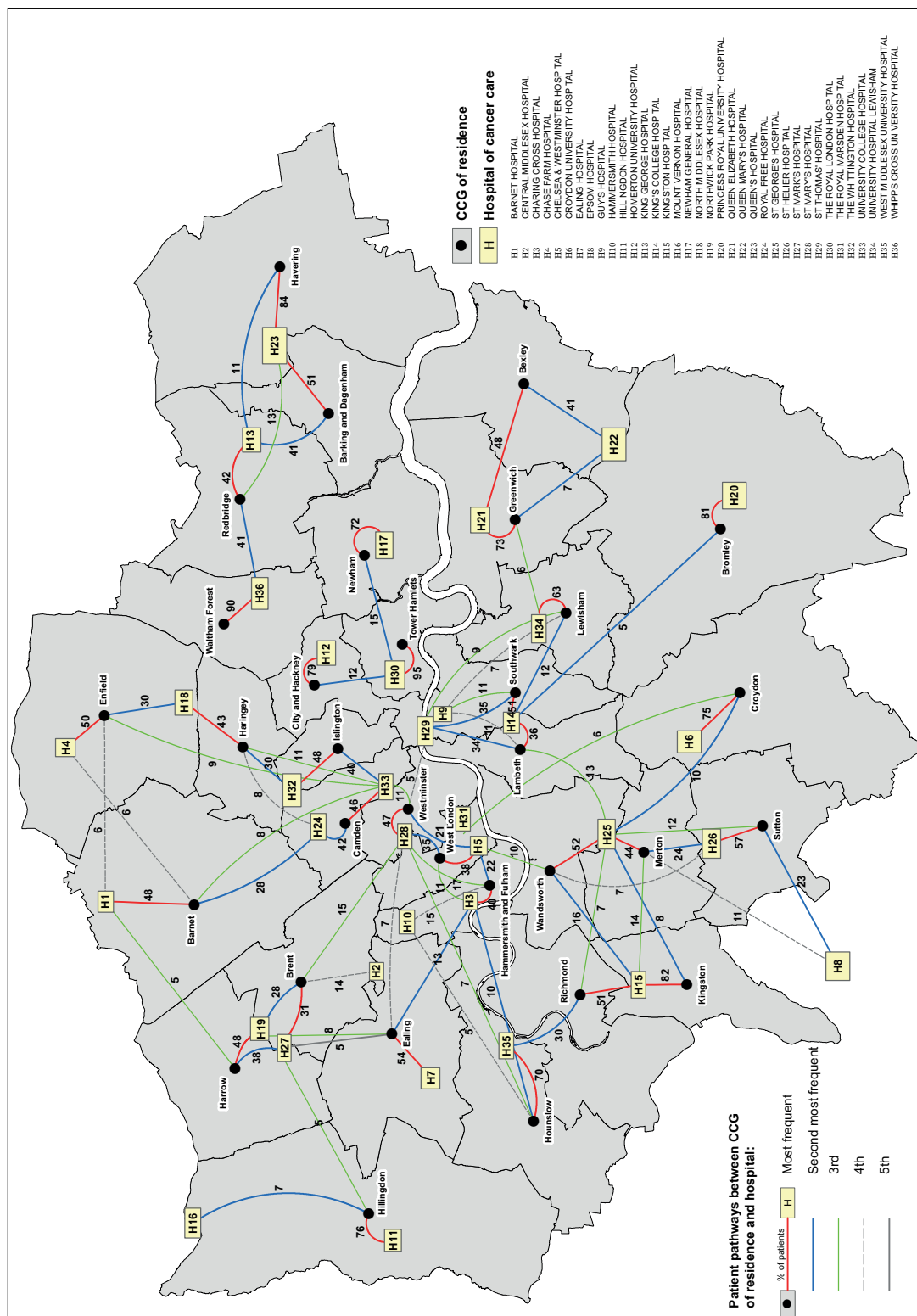


Figure 6.3: Flow map of London displaying the pathways of patients' journeys between the CCG of residence and the hospital of cancer care for men diagnosed with colon cancer, 2006-2013.

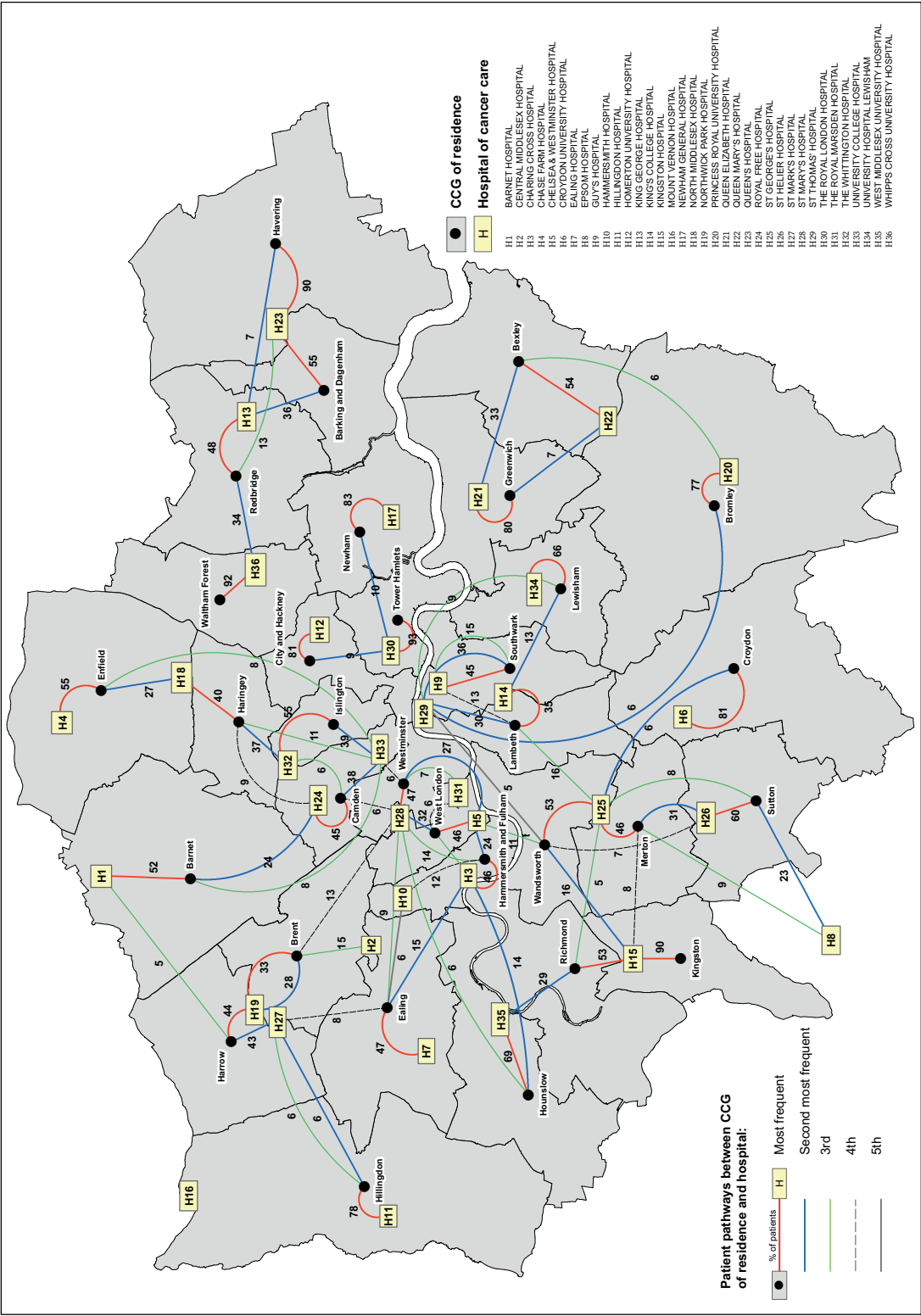


Figure 6.4: Flow map of London displaying the pathways of patients' journeys between the CCG of residence and the hospital of cancer care for women diagnosed with colon cancer, 2006-2013.

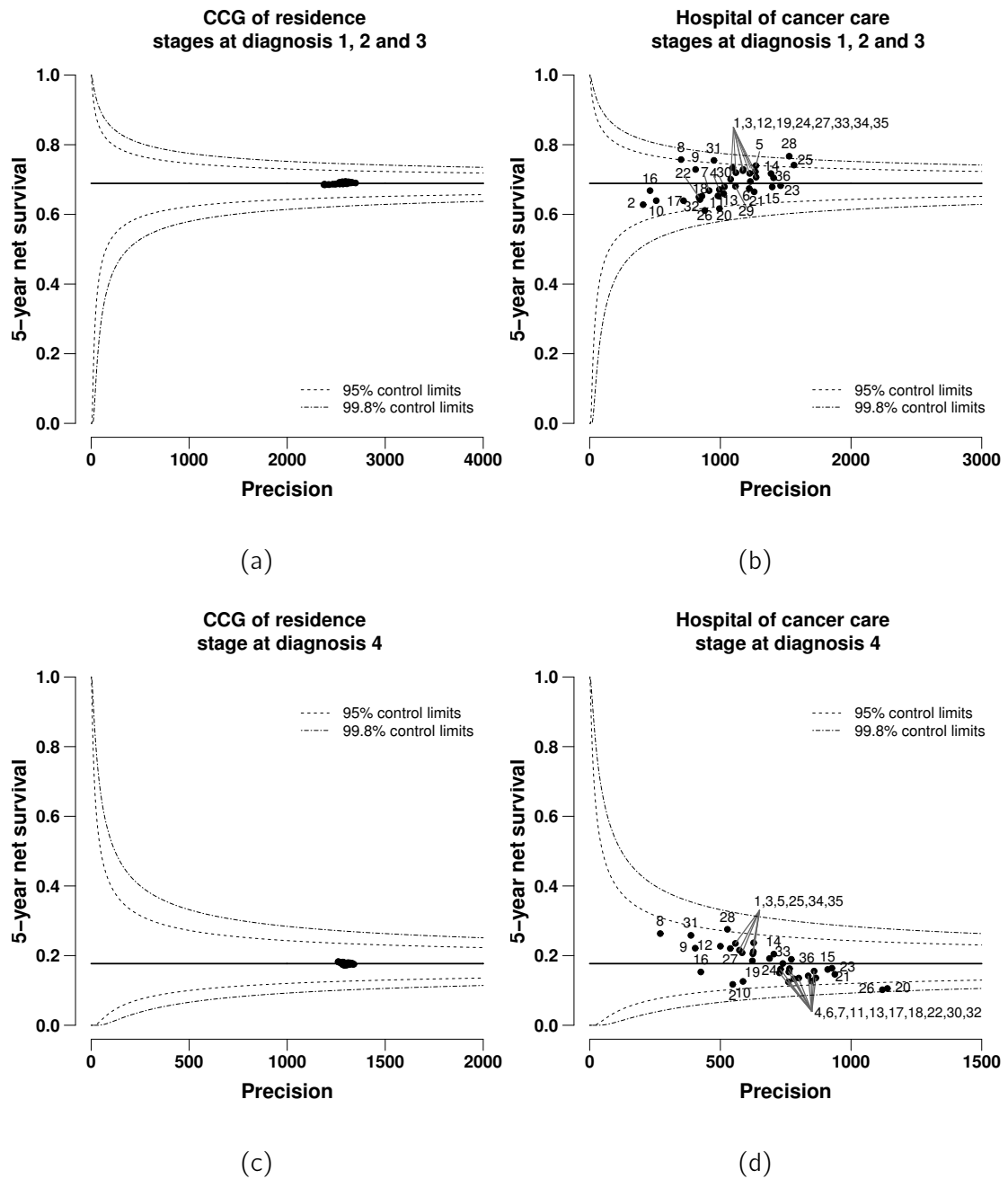


Figure 6.5: Funnel plots of 5-year net survival (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) by CCG of residence for stages at diagnosis 1, 2 and 3; (b) by hospital of cancer care for stages at diagnosis 1, 2 and 3; (c) by CCG of residence for stage at diagnosis 4; (d) by hospital of cancer care for stage at diagnosis 4.

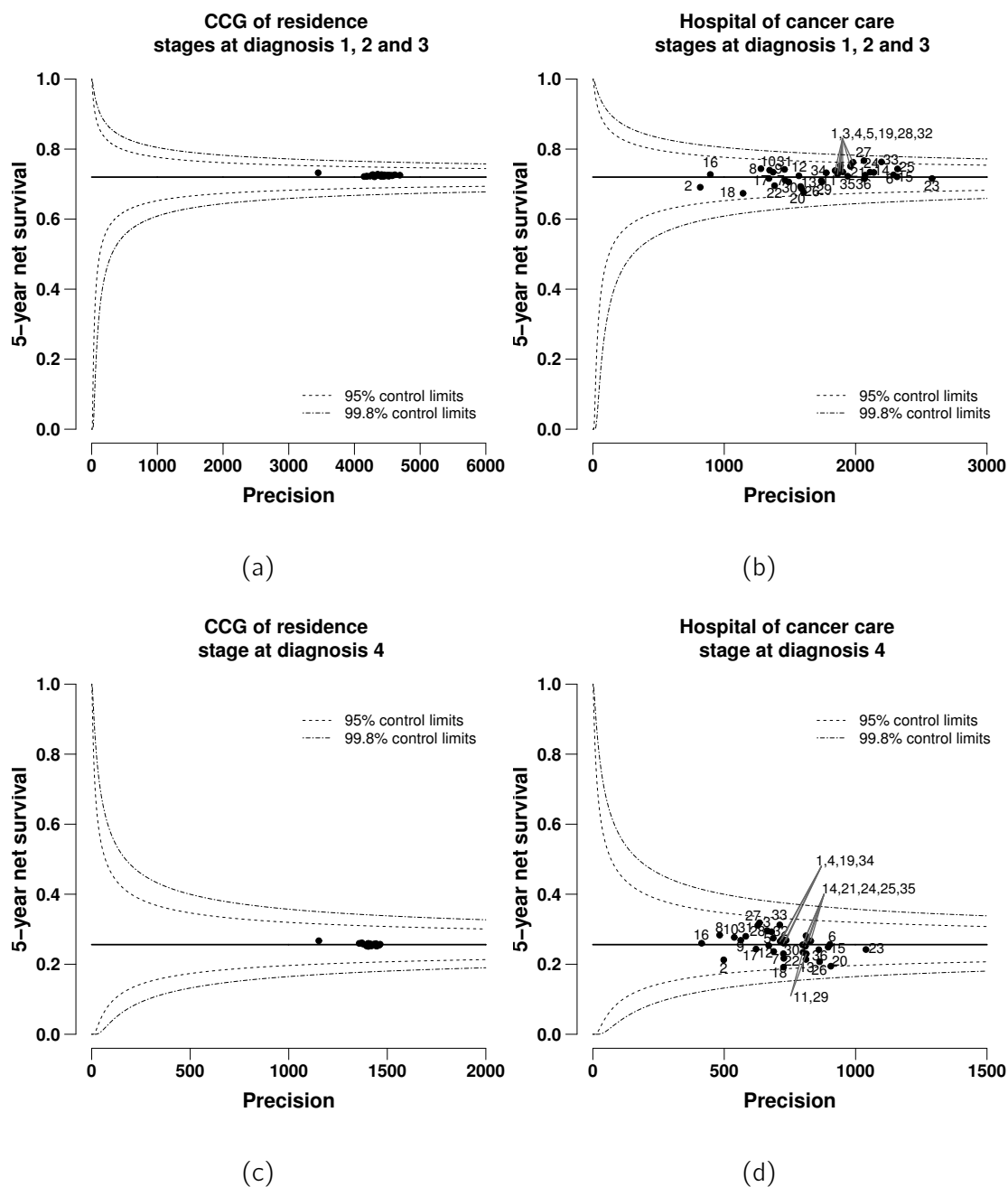


Figure 6.6: Funnel plots of 5-year net survival (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) by CCG of residence for stages at diagnosis 1, 2 and 3; (b) by hospital of cancer care for stages at diagnosis 1, 2 and 3; (c) by CCG of residence for stage at diagnosis 4; (d) by hospital of cancer care for stage at diagnosis 4.

Discussion

Cancer survival studies in England have mostly focussed on describing geographical disparities, with strong evidence suggesting that the place of residence plays an important role in the survival of cancer patients. In this article we hypothesised whether the observed disparities in cancer survival were more associated with the place of residence or with the place (hospital) where cancer patients receive care: ‘What matters more, where you live or where you receive care?’. We investigated variation in colon cancer survival for patients living and receiving care in London between 2006-2013. Flow maps of patient pathways between the area of residence and the hospital of cancer care revealed that patients travelled more frequently to hospitals closest to their area of residence, mainly in the North East, East and South East of London. Whereas patients living in the North and South West of London also frequently accessed hospitals outside their area of residence. Wide variation was observed in five-year net survival between CCGs, even after adjusting for age at diagnosis, deprivation and stage at diagnosis. These disparities reduced once adjusted for hospital of cancer care, while hospital variation remained even after adjusting for patient- and tumour-level characteristics; However, there is a strong correlation between CCGs and hospitals as patients tend to go to the nearest hospital and this pattern is particularly strong in some of the most deprived London CCGs.

CCGs have the responsibility to allow patients to make choices and to promote their involvement in decisions related to their care or treatment [258]. The differential frequencies in patient pathways between area of residence and hospital of care raise questions regarding the equal choice of patients for the best performing hospitals at point of referral.

To the best of our knowledge this is the first study to investigate variation in cancer survival at both CCG and hospital level. We advocate caution when interpreting the hospital-specific net survival estimates presented in this study. These levels of survival cannot be imputed to any individual hospital included in this study since these hospitals treat more patients than the selected cohort of cancer patients here analysed. The survival variations observed relate solely to this cohort of patients and cannot be generalised to all the patients seen in each hospital.

Stage at diagnosis was not available for 23% of the cases. In order to include all the cases in the analysis, we extended the excess hazard model by specifying an additional distribution for the stage variable (regardless if observed or not) that uses information from all the covariates included in the main model specification. Additional analysis performed on complete cases confirmed the practical importance and the impact on results of accommodating the missing data structure in the analysis (see results in Appendix A6).

In summary, this study demonstrates the importance of performing more in-depth investigations into the observed disparities in cancer survival using population-based data enriched with other relevant health data sources. Future work should aim to investigate hospitals with poorer performance to understand its causes, including resources and organisation among other factors. And to examine more in depth (including qualitative studies) what determines the choice (or absence of choice) of patients for a given of hospital in order to suggest actions to correct such wide disparities.

Appendix A. Flexible Bayesian hierarchical excess hazard models

Appendix A.1. Model specification

Excess hazard models were set-up for men and women, including age at diagnosis (AGE), deprivation category (DEP), stage at diagnosis (STAGE), CCG of residence (CCG) and hospital of care (HOSP). The models were defined on the log-excess hazard scale and use low-rank thin plate (LRTP) splines to model the smooth effect of the baseline excess hazard and the smooth effect of age at diagnosis (6). The observed follow-up time (t) was divided into four partitions ($K=4$), chosen at the 25%, 50% and 75% percentiles of the event (death) times. For men these were chosen at $\tilde{t}=(0, 0.28, 1.08, 2.4, 8)$ years and for women at $\tilde{t}=(0, 0.27, 1, 2.3, 8)$ years. Both models, for men and women, were formulated as

$$\begin{aligned}
 \log(h_E(t|\alpha;\beta;\gamma;\nu;\iota;\zeta)) = & (\alpha_{0,0} + \alpha_{1,0}AGE) + (\alpha_{0,1} + \alpha_{1,1}AGE)t \\
 & + \sum_{k=2}^K (\alpha_{0,k} + \alpha_{1,k}AGE)(|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \quad [\text{part 1}] \\
 & + \beta_1^*(AGE - \overline{AGE}) + \sum_{j=2}^J \beta_j^*(|AGE - \widetilde{AGE}_{j-1}|^3 \\
 & - |\overline{AGE} - \widetilde{AGE}_{j-1}|^3) \quad [\text{part 2}] \\
 & + \sum_{l=2}^5 (\gamma_l * DEP_l) \quad [\text{part 3}] \\
 & + \nu * STAGE \quad [\text{part 4}] \\
 & + \sum_{v=1}^{32} (\iota_v * CCG_v) \quad [\text{part 5}] \\
 & + \sum_{h=1}^{36} (\zeta_h * HOSP_h) \quad [\text{part 6}]
 \end{aligned} \tag{6.1}$$

where, [part 1] formulates the LRTP spline modelling the baseline log-excess hazard, incorporating the time-dependent effect of age at diagnosis using the same follow-up time partition, with parameters $\alpha = (\alpha_0|\alpha_1)$ and $\alpha_q = (\alpha_{q,0}, \dots, \alpha_{q,K})$ for $q=0,1$. [part 2] represents the LRTP spline modelling the non-linear (smooth) effect of age at diagnosis using 3 partitions ($J=3$) of the observed age range at $\widetilde{AGE}=(15, 43, 71, 99)$ years, for

both men and women, with parameters β_j , $j = 1, \dots, J$. \overline{AGE} represents the mean age at diagnosis. For ease of interpretation, age at diagnosis was centered at age 70. [part 3] formulates the effect of deprivation modelled as a categorical variable (DEP_1 : least deprived to DEP_5 : most deprived), with parameters γ_l , $l = 2, \dots, 5$. The least deprived group (DEP_1) was set as the baseline fixing $\gamma_1 = 0$. [part 4] formulates the effect of stage at diagnosis modelled as a binary variable ($STAGE=0$ for stages 1, 2 and 3 grouped and $STAGE=1$ for stage 4), with parameter ν . [part 5] defines the random effects for CCG of residence, with parameters ι_v , $v = 1, \dots, 32$. [part 6] defines the random effects for hospital of care, with parameters ζ_h , $h = 1, \dots, 36$.

Appendix A.2. Prior distributions

Prior distributions for the model parameters were chosen as:

- For the baseline log-excess hazard, including the time dependent effect of age at diagnosis ([part1]):

$$\begin{aligned} \alpha_{q,0} &\sim N(0, 10^4), \alpha_{q,1} \sim N(0, 10^4) \text{ for } q=0, 1 \\ \alpha_{q,k} | \sigma_{q,\alpha} &\overset{iid}{\sim} N(0, \sigma_{q,\alpha}^2) \text{ for } k=2, \dots, K \text{ and } \sigma_{q,\alpha} \sim U(0.01, 100) \text{ for } q=0, 1 \end{aligned} \quad (6.2)$$

- For the non-linear effect of age at diagnosis ([part 2]):

$$\begin{aligned} \beta_0 &\sim N(0, 10^4) \\ \beta_k | \sigma_\beta &\overset{iid}{\sim} N(0, \sigma_\beta^2), \text{ for } k=2, \dots, K \text{ and } \sigma_\beta \sim U(0.01, 100) \end{aligned} \quad (6.3)$$

- For the effect of deprivation ([part 3]):

$$\begin{aligned} \gamma_0 &= 0 \\ \gamma_l | \sigma_\gamma &\overset{iid}{\sim} N(0, \sigma_\gamma^2), \text{ for } l=2, \dots, 5 \text{ and } \sigma_\gamma \sim U(0.01, 100) \end{aligned} \quad (6.4)$$

- For the effect of stage at diagnosis ([part 4]):

$$\nu \sim N(0, 10^4) \quad (6.5)$$

- For the random effects on CCG of residence ([part 5]):

$$\iota_v | \sigma_\iota \stackrel{iid}{\sim} N(0, \sigma_\iota^2), \text{ for } v=1, \dots, 32 \text{ and } \sigma_\iota \sim U(0, 10) \quad (6.6)$$

- For the random effects on hospital of cancer care ([part 6]):

$$\zeta_h | \sigma_\zeta \stackrel{iid}{\sim} N(0, \sigma_\zeta^2), \text{ for } h=1, \dots, 36 \text{ and } \sigma_\zeta \sim U(0, 10) \quad (6.7)$$

Appendix A.3. Handling missing information on stage at diagnosis

Information on stage at diagnosis was missing for 22% of men and 24% of women in the dataset analysed in this study. All other variables had no missing information. In order to include all the cases in the analysis, we extended the model specified in 6.5 to define a prior distribution for stage at diagnosis using a Bernoulli distribution with probability μ as

$$STAGE \sim \text{Bernoulli}(\mu) \quad (6.8)$$

and we defined as a prior distribution for μ a logistic regression model including all the covariates used in the main model to better impute the missing stage information as

$$\begin{aligned} \text{logit}(\mu) = & \lambda_1 * AGE_i + \sum_{l=2}^5 (\lambda_{2l} * DEP_l) \\ & + \sum_{v=1}^{32} (\lambda_{3v} * CCG_v) + \sum_{h=1}^{36} (\lambda_{4h} * HOSP_h) \end{aligned} \quad (6.9)$$

where, AGE_i is now modelled as a linear effect of age at diagnosis, with parameter λ_1 . The effects of deprivation with parameters λ_{2l} , of CCGs with parameters λ_{3v} and of hospitals with parameters λ_{4h} are modelled in the same way as in the main model formulation (6.5). Prior distributions for all the λ parameters were defined as

$$\begin{aligned}
\lambda_1 &\sim N(0, 0.0001) \\
\lambda_{2l} &\overset{iid}{\sim} N(0, 0.0001) , \text{ for } l=2, \dots, 5 \\
\lambda_{3v} &\overset{iid}{\sim} N(0, 0.0001) , \text{ for } v=1, \dots, 32 \\
\lambda_{4h} &\overset{iid}{\sim} N(0, 0.0001) , \text{ for } h=1, \dots, 36
\end{aligned} \tag{6.10}$$

Appendix A.4. Bayesian inference

Bayesian inferences were performed in R software version 3.4.3 using the JAGS MCMC program accessed via the R package 'R2JAGS' (10,11). Models were fitted setting up 2 MCMC chains, each with 60,000 iterations, a burn-in period of 10,000 and a thinning of 2 to eliminate any potential autocorrelation among samples within the chains. A total of 50,000 sampled values were retained from the posterior distributions of each of the model parameters. An examination of the trace and density plots of each parameter's posterior distribution did not indicate any convergence issues for these samples. The 50,000 sampled values from the parameter posterior distributions were used to derive posterior distributions of 5-year net survival for each CCG of residence and hospital of care. These were derived using a 'prediction matrix' that included all the combinations of age at diagnosis (individual integer ages within the observed age range 15-99 years), deprivation category (1-5), stage at diagnosis (0-1), CCG (32) and hospital (36).

Appendix A.5. Funnel plots for the additional models fitted

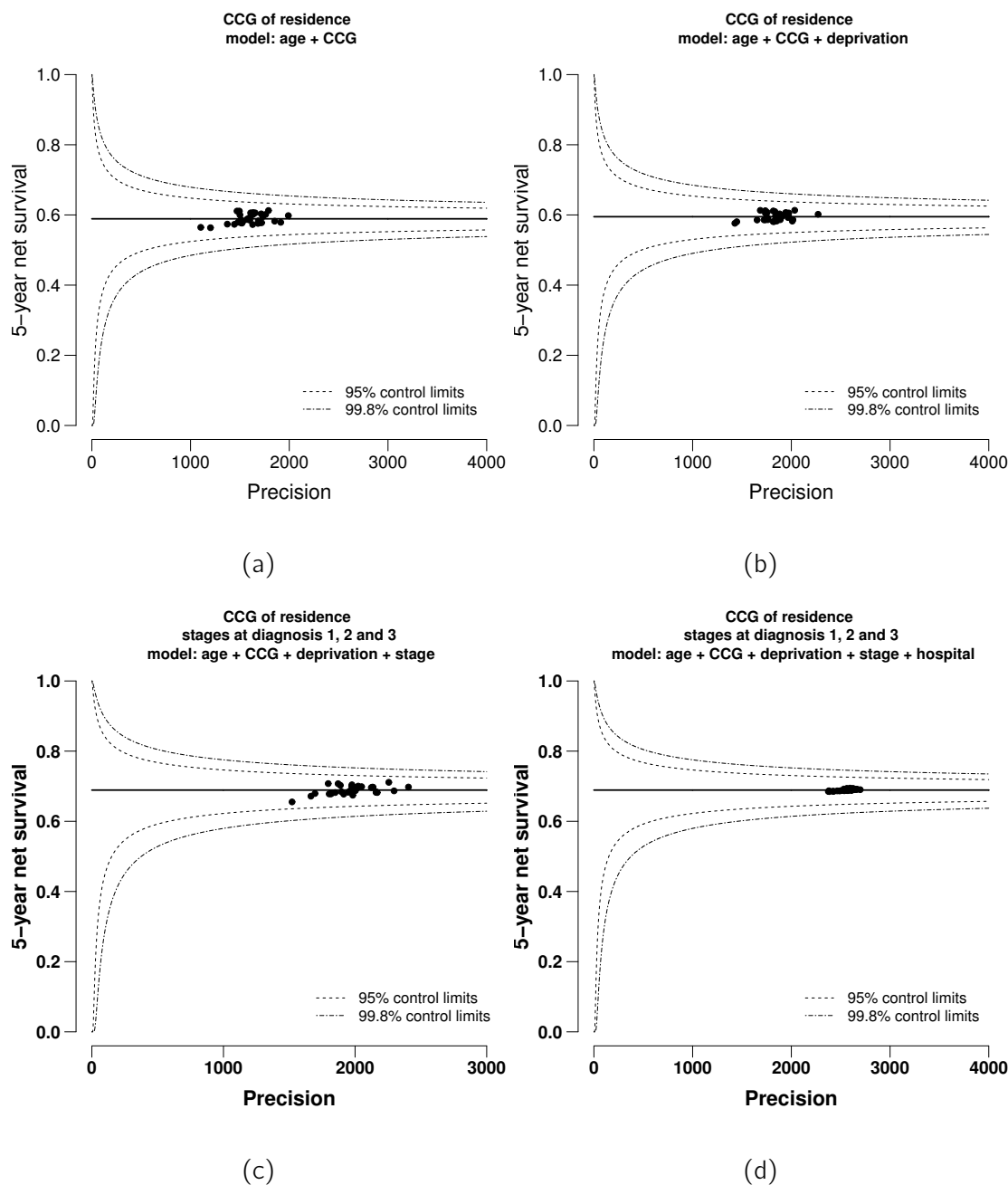


Figure 6.7: Funnel plots of 5-year net survival by CCG of residence (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stages at diagnosis 1, 2 and 3); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stages at diagnosis 1, 2 and 3).

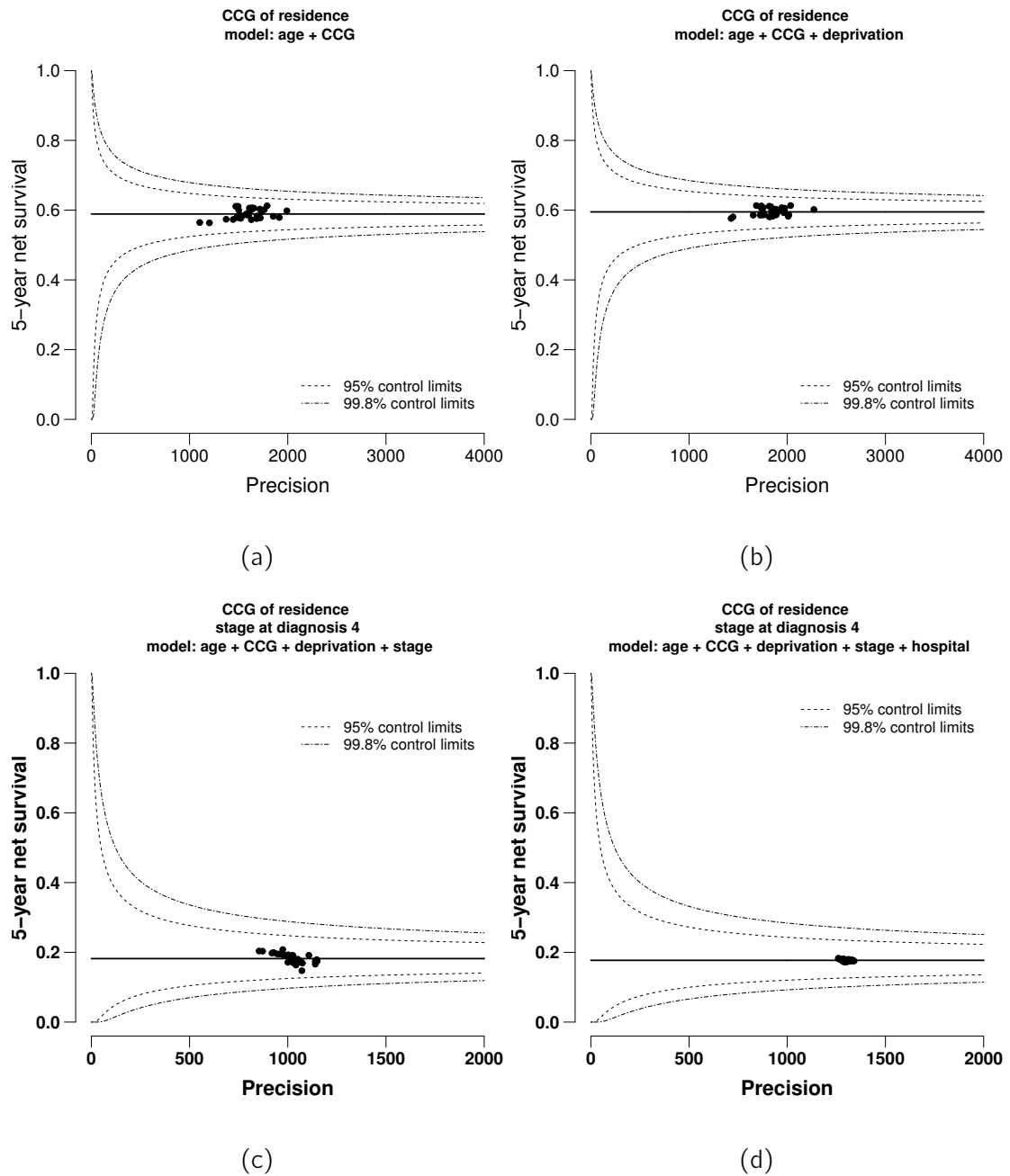


Figure 6.8: Funnel plots of 5-year net survival by CCG of residence (mean posterior) for men diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stage at diagnosis 4); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stage at diagnosis 4).

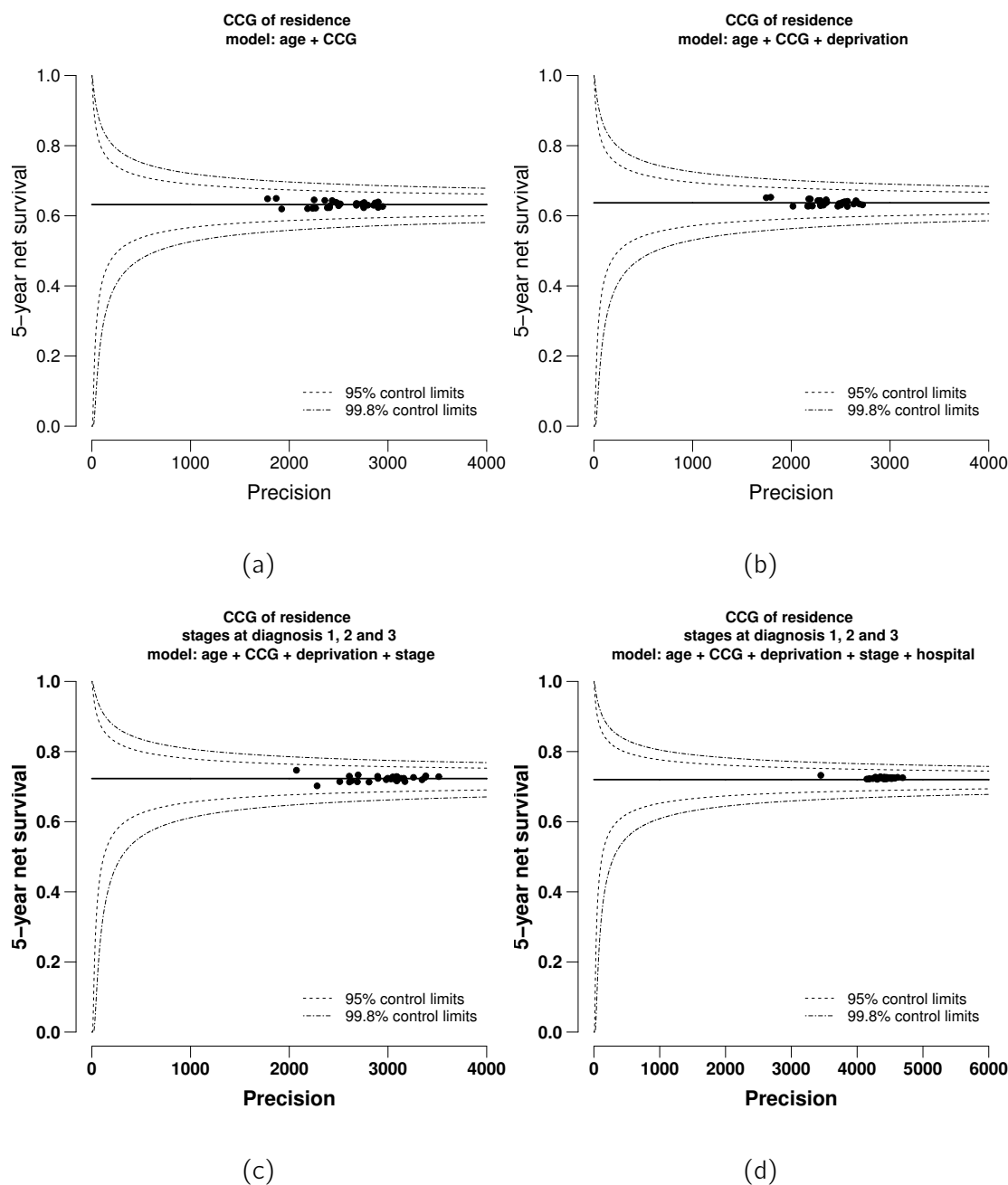


Figure 6.9: Funnel plots of 5-year net survival by CCG of residence (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stages at diagnosis 1, 2 and 3); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stages at diagnosis 1, 2 and 3).

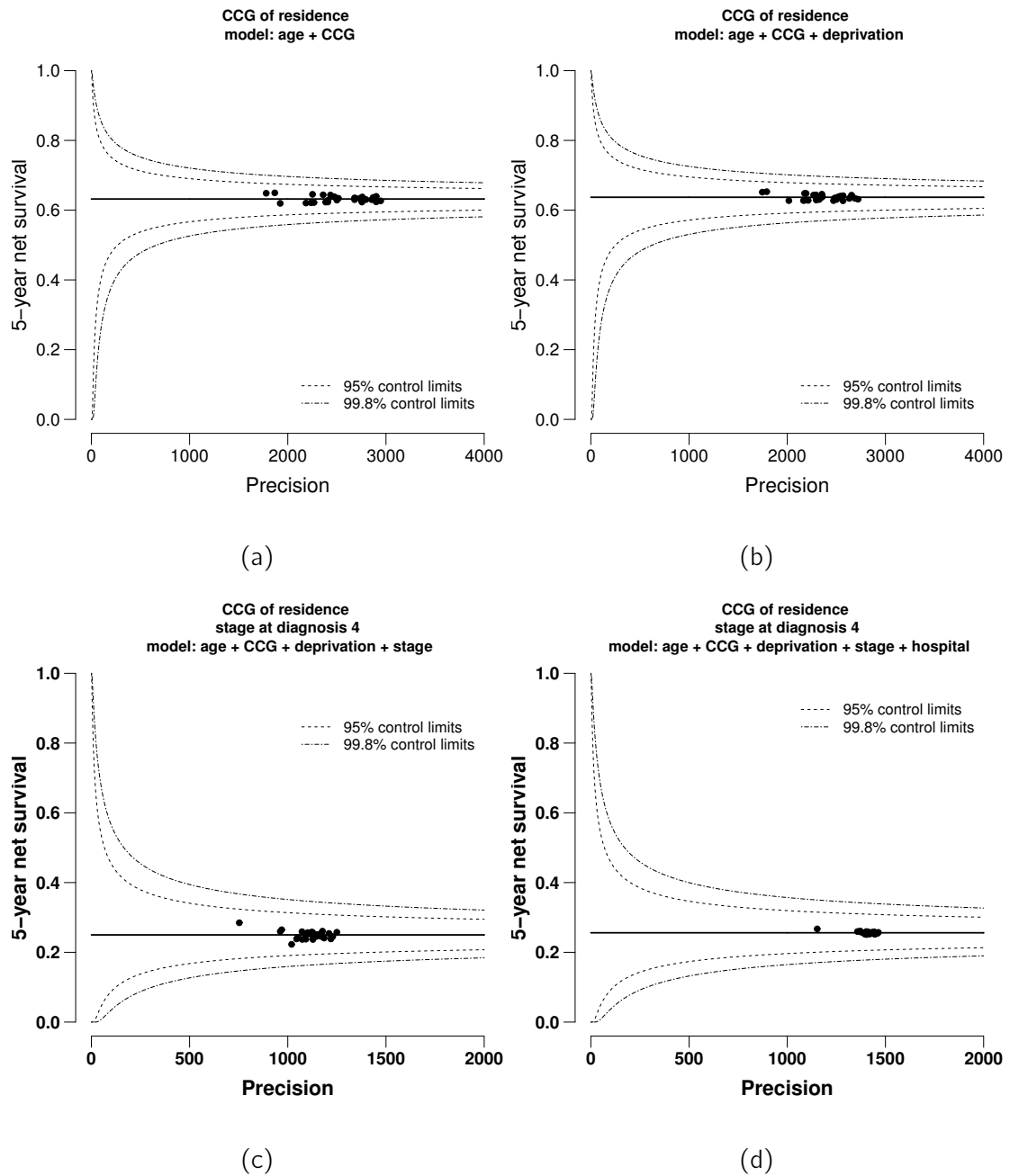


Figure 6.10: Funnel plots of 5-year net survival by CCG of residence (mean posterior) for women diagnosed with colon cancer in 2006-2013, London: (a) model including age at diagnosis and CCG (b) model including age at diagnosis, CCG and deprivation; (c) model including age at diagnosis, CCG, deprivation and stage at diagnosis (for stage at diagnosis 4); (d) model including age at diagnosis, CCG, deprivation, stage at diagnosis and hospital of cancer care (for stage at diagnosis 4).

Appendix A.6. Funnel plots for the complete case analysis versus modelling missing data

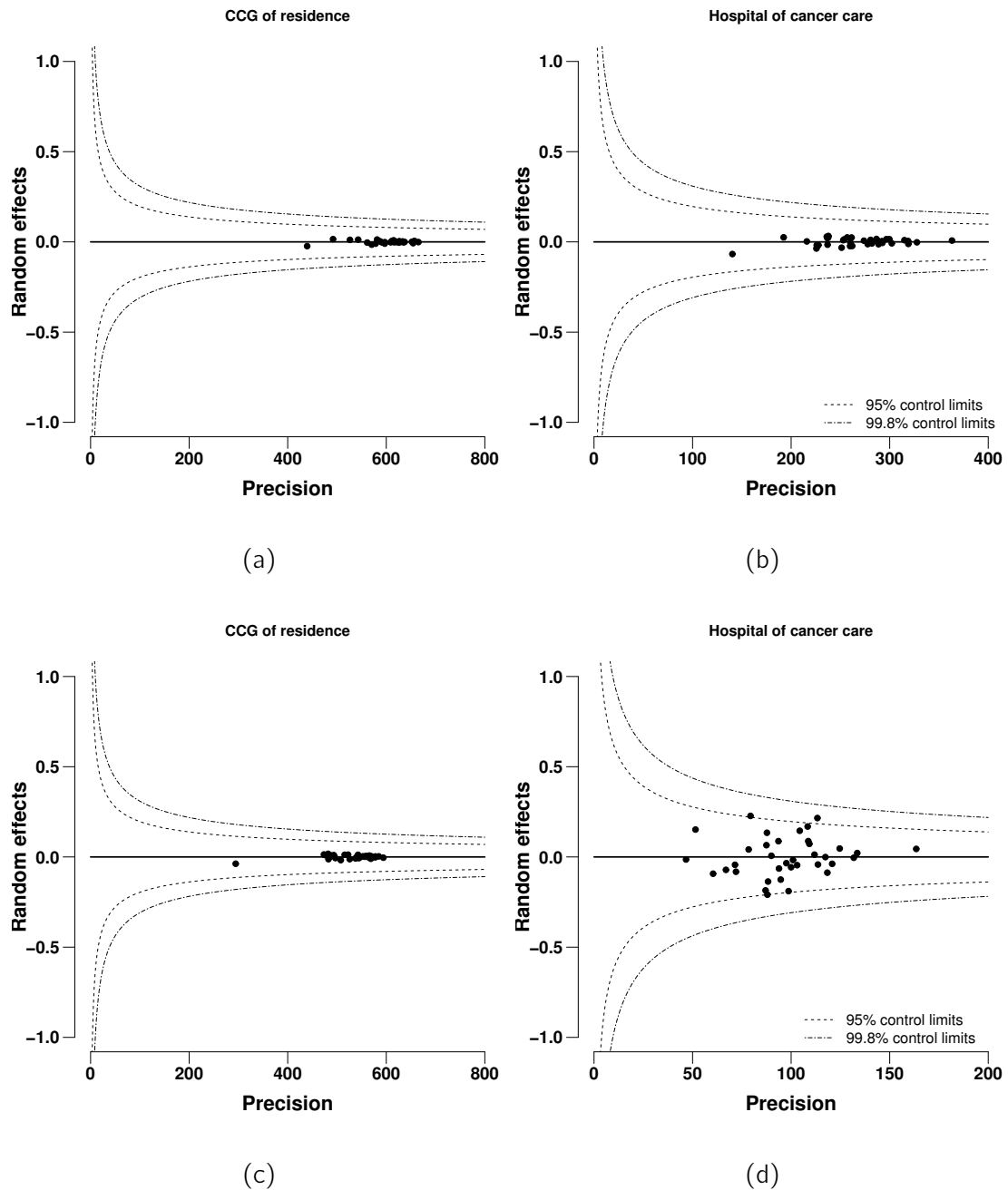


Figure 6.11: Funnel plots of the random effects by CCG of residence and hospital of care for women using: complete case analysis after removing cases with missing stage at diagnosis ((a) and (b)) and using all data by modelling the missing data structure ((c) and (d)).

References

- (1) Shack, L.G., 2009. What factors influence socio-economic inequalities in colorectal cancer survival? Ph.D. thesis. London School of Hygiene and Tropical Medicine.
- (2) Benitez Majano, S., Di Girolamo, C., Rachet, B., Maringe, C., Guren, M., Glimelius, B., Iversen, L., Schnell, E., Lundqvist, K., Christensen, J., Morris, M., Coleman, M.P., Walters, S. 2019. Surgical treatment and survival from colorectal cancer in denmark, england, norway and sweden: A population-based study. *Lancet Oncology* 20(1): 74-87.
- (3) Fowler, H., Belot, A., Njagi, E.N., Luque-Fernandez, M.A., Maringe, C., Quaresma, M., Kajiwarra, M., Rachet, B. 2017. Persistent inequalities in 90-day colon cancer mortality: an english cohort study. *British Journal of Cancer* 117: 1396-1404.
- (4) STATA statistical software: Release 15. 2017. StataCorp LLC. URL: <https://www.stata.com/>.
- (5) ArcGIS software: Release 10.5. 2017. ESRI. URL: <https://www.arcgis.com/>.
- (6) Quaresma, M., Carpenter, J., Rachet, B. 2019. Flexible bayesian excess hazard models using low-rank thin plate splines. *Statistical Methods in Medical Research*.
- (7) Rachet, B., Maringe, C., Woods, L.M., Ellis, L., Spika, D., Allemani, C. 2015. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health* 15: 1240.
- (8) Cancer Survival Group. 2019. Life tables for England and Wales by sex, calendar period, region and deprivation. London School of Hygiene and Tropical Medicine.
- (9) Quaresma, M., Coleman, M.P., Rachet, B. 2013. Funnel plots for population-based cancer survival: principles, methods and applications. *Statistics in Medicine*.
- (10) RStudio Team. RStudio: Integrated Development for R. 2015. URL <http://www.rstudio.com/>.
- (11) Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.

6.6 Discussion

In this last research chapter we aimed to determine how Bayesian approaches could be used in the relative survival setting to improve the estimation of cancer survival in the presence of sparse data, and when using more complex data structures, including hierarchical and spatially arranged data.

A summary of the existing literature on small area estimation methods revealed that very few models are available in the relative survival setting. In the absence of suitable Bayesian regression models defined on the log-excess hazard scale that could model complex data structures, i.e. extending the Estève et al. [81] model for Bayesian inference, we defined a set of characteristics that a new model should satisfy. Based on these characteristics, we proposed a flexible Bayesian excess hazard model on the log-excess hazard scale based on the full-likelihood specification using individual-level data. We chose to use Low-Rank Thin Plate splines (LRTP splines) to model the various components of the excess hazard model because these splines offer a reasonable compromise between model flexibility and likelihood tractability. As discussed in Research Publication 3, these splines are simple yet flexible first-order polynomials that provide a good alternative to other spline constructs commonly used in excess hazard models, as for instance restricted cubic splines mentioned in Chapter 2. For models fitted on the log-excess hazard scale using those other commonly used splines will very frequently add complexity to the likelihood specification, requiring numerical integration techniques to evaluate it [99, 259]. Models fitted on the log cumulative excess hazard scale have the advantage of avoiding the use of numerical integration because of the resulting tractable cumulative excess hazard and excess hazard functions, but the interpretation of multiple time-dependent effects is difficult because the excess hazard ratio for one variable can depend on the levels of the other variables, even without having defined interaction terms in the model [97].

Along with the new flexible model formulation we also implemented a post-estimation procedure to derive posterior distributions for the excess hazard ratios, excess hazard functions and net survival based on the saved MCMC samples for each parameter. One downside of fitting this model is that it can be computationally very intensive, varying from a few hours to a few days to complete the computation, depending on the model complexity,

the number of MCMC iterations and the size of the matrices generated. Although computation time can be reduced by the use of parallel computing, some further improvements are needed to reduce computation time and improve sampling performance to model more complex data structures.

The work presented in this last chapter lays out the foundation model for flexible excess hazard modelling within the Bayesian framework. One of the characteristics that we set up a priori for the implementation of the new excess hazard model was the easy extension to accommodate hierarchical and spatial data structures. When reviewing the literature, most incidence and survival studies were very focussed on the idea of defining a spatial structure based on borrowing strength from neighbouring areas. We hypothesise that in the case of cancer survival, it does not matter so much the area where patients live but the care facility where they are treated. As a first extension of the flexible Bayesian model, we included a pair of random effects to investigate variation in colon cancer survival in London at the CCG and hospital of care level. The main results showed that after adjusting for age at diagnosis, socioeconomic status, stage at diagnosis and hospital of care, no variability in survival was observed between CCGs, contrasting with a much more pronounced variability between hospitals. Taking into account these results, future work should focus on how to define the most appropriate neighbouring structure for cancer survival analysis. This structure will ideally not be based on a neighbouring area that shares a common border or purely based on distance, but we would like to identify areas that share common traits, such as common management or treatment protocols and translate that information into a dependency structure to incorporate into the flexible Bayesian excess hazard model.

Chapter 7

Discussion and Conclusions

The research I present in my doctoral thesis is divided into three interconnected Research Aims that arose as a natural research progression to further and complement each research output. The initial research idea (Research Aim 1) originated from a request by national policy-makers to provide ‘one single’ number that could summarise the patterns of survival for all cancers combined in England. This ‘summary number’ was envisioned to act as a simple and informative monitoring tool for cancer survival at both national and local level. At national level, to act as a surveillance tool of strategic value and at local level, to act as a monitoring tool for local health service managers. When faced with the request of providing one summary survival number for all cancers combined, it was clear that this number could not be a simple survival average of all cancer types pooled together. I created the concept of the cancer survival index to ensure that differential distributions of cancer patients by sex, age and cancer type, or any shifts in these distributions over time do not influence comparisons between populations. It is the first index of this type to be introduced in England designed specifically to aid health-policy makers and healthcare managers monitor and assess the effectiveness of cancer survival outcomes. However, caution is required in its interpretation. The index does not reflect the prospects of survival for any individual cancer patient. It should not be interpreted as the only indicator of performance, but in conjunction with other information available for that country or region. It should be seen as a guide to raise questions about the potential for improvement.

The two applications I present for the England and the CCG survival indexes, demonstrate that the index can be used at any level of geographical aggregation. Although the concept of the survival index is simple, the estimation process, in particular the modelling strategy are complex. To overcome estimation challenges, I developed a 'semi-automated' modelling strategy that made it feasible to estimate the many individual components needed for the construction of the indexes. It has also provided a more robust estimation of those individual components through the modelling of age and year of diagnosis. However, the estimation of survival remained challenging in some situations, in particular for some of the smallest CCGs, and the CCG index did not prove feasible to estimate beyond one-year since diagnosis. Improvements to the modelling strategy can be made when estimating the indexes for smaller geographies as discussed in Chapter 4, or through the use of other modelling approaches such as the ones proposed as part of Research Aim 3, which I will discuss below.

The approach I propose to construct the cancer survival index is timeless and can be applied to other health geographies using the same set of sex-age-cancer specific weights. The novel modelling strategy, developed to improve the estimation of the individual index components, was presented in a detailed way to facilitate and guide other researchers interested in developing a cancer survival index for their setting. This work has already motivated other countries to construct their own cancer survival indexes using the same approach. The United States constructed a North American Cancer Survival Index to Measure Progress of Cancer Control Efforts [160] and Japan started to develop a national index of cancer survival (work in progress).

Since publication, both the national and the local cancer survival indexes have attracted much attention [1–5]. The results for the national cancer survival index supported CRUK's vision set out in their 2014 research strategy [155]: *'Cancer Research UK's vision is to bring forward the day when all cancers are cured. Over the last 40 years, cancer survival rates in the UK have doubled. In the 1970s just a quarter of people survived. Today that figure is half. Our ambition is to accelerate progress and see three-quarters of patients surviving the disease within the next 20 years.'* The same results have been fed into numerous CRUK's public funding campaigns and into online information blogs [156].

However, presenting the results of the cancer survival index to 'lay' or non-technical audiences revealed a new challenge in particular when presenting the results for the CCG survival index. Very long tables of results or 'standard' bar charts packed with many bars make it very difficult to communicate any observed patterns in a meaningful way. After some research, I came across two data visualisation techniques used in other areas (smoothed maps for incidence outcomes and funnel plots for mortality outcomes), and I decided to adapt these two techniques to cancer survival outcomes to improve the visualisation of cancer survival for a more successful dissemination of these outcomes to policy-makers. This work led to the development of Research Aim 2.

These visualisation tools were successfully used in diverse contexts. For example, the publication of the CCG cancer survival index gave me the opportunity to present the results at the annual meeting of the All-Party Parliamentary Group on Cancer (APPGC), chaired by John Baron MP. APPGCs are cross-party groups run by and for Members of the Commons and Lords to discuss a range of relevant topics. Only 10 minutes were allocated to the presentation and discussion of results. The presentation had to be quick and clear for the main message to be successfully delivered. Only smoothed maps and funnel plots were used to show the trends in the one-year cancer survival index by CCGs (the presentation slides are provided in Appendix B). The outcome of the meeting was transformational. Soon after, APPGC advocated that one of their main aims was to ensure that local commissioners were held accountable for improving one-year survival in their areas. APPGC worked to persuade the top tiers of the NHS to include cancer survival in the Delivery Dashboard. The decision was announced by the Chief Executive of NHS England, Simon Stevens, when addressing the 'Britain Against Cancer' conference in December 2015, that the CCG cancer survival index was to be included in the Delivery Dashboard of the NHS' Assurance Framework, that this sits at the top of the NHS accountability tree [157–159].

Presenting the survival index using only these two data visualisation techniques was decisive for the success of the APPGC meeting. Smoothed maps and funnel plots proved to be two simple and intuitive data visualisation tools. These were in particular powerful in a presentation mode showing animations of maps and funnel plots looping over several years of diagnosis for easy visualisation of the time trends in the survival indexes.

Since publication, smoothed maps and funnel plots have been used by the Department of Health for overall strategy purpose and for local management [177]. Funnel plots have also been used to display regional and race specific variation in population-based cancer survival in the United States [261]. These two data visualisation techniques also led to the development of a lecture I give on data visualisation techniques for cancer survival outcomes for the short course 'Cancer Survival: Principles, Methods and Applications' held annually at LSHTM.

The last research aim of my thesis (Research Aim 3), addressed the estimation challenges faced in Chapter 4 for the estimation of the individual components of the indexes. For this purpose, I decided to explore how Bayesian approaches could be used in the relative survival setting to improve the estimation of cancer survival in the presence of sparse data, and when using more complex data structures, including spatially arranged and hierarchical data. When summarising the literature for small-area estimation I only found a few models for the estimation of excess hazards within the Bayesian framework, and none for the estimation of net survival. As an initial step, I proposed a flexible Bayesian excess hazard model on the log-scale based on the full-likelihood specification and provided a step-by-step tutorial on the estimation of net survival from this model. The model uses low-rank thin plate splines providing a compromise between model flexibility and likelihood tractability, specially important within the Bayesian framework and then modelling more complex data structures as it reduces the computational burden. The benefits of using this type of splines was demonstrated in the subsequent application that extended the model to include two random effects to investigate cancer survival variation in London.

The field of population-based cancer survival continues to be a very active area of methodological research. In light of the work presented in this thesis, I suggest as further lines of research: a) To implement the alternative modelling strategy described in the Discussion of Chapter 4 to further improve the estimation of the individual components of the index. An additional avenue of research for the estimation of the index is to explore the joint modelling of two or more cancers of similar survival patterns to improve prediction of the individual index components. b) The research presented on funnel plots for cancer survival has already led to the development of further research. Additional work has been

done to provide a set of guidelines to handle over-dispersion in cancer survival outcomes (manuscript in review). c) To extend the flexible Bayesian excess hazard model by incorporating the most adequate spatial dependency structure enabling excess hazard spatial regression to be performed. To implement a dedicated MCMC sampler to improve computational speed and sampling performance when using the flexible Bayesian excess hazard model with more complex data structures.

In conclusion, cancer survival is the metric of choice when assessing and monitoring the effectiveness of healthcare systems in treating and caring for cancer patients. Persistent inequalities in cancer survival have been reported for England over the last 5 decades, with lower survival typically observed in the North of the country. Since the mid 1990's, large survival disparities have also been reported between England and countries considered to be of equivalent wealth and similar healthcare systems, with lower survival observed in England. These survival deficits have led to many health-policy related initiatives aimed at tackling cancer inequalities and achieving world-class cancer survival outcomes for England [68, 262]. Monitoring progress in cancer survival over time became essential to assure that these objectives are met, with many survival outcome indicators being published on a regular basis for different levels of geographical aggregation in England. However, the constant changes to the configurations of these health geographies are reflected in increased pressures and demands from health policy-makers and healthcare managers for the timely availability of monitoring tools for the changing health structures. Investigating cancer survival inequalities and disentangling the factors that might contribute to these discrepancies is a complex task, and many avenues of research can be undertaken. The research I present in this thesis focused on a few methodological aspects to improve the estimation and dissemination of cancer survival to a vast range of audiences. As the research questions become more complex, more robust methods and more detailed quality cancer data are important, but it is also primordial that our results are useful and accessible to policy makers. In the era of electronic health records and the existence of many rich and complementary sources of data, it is imperative to guarantee that researchers continue to have timely access to all the sources of cancer data for the continued success of population-based cancer research.

Appendix A

Stata and R code

A.1 Stata code to estimate the national and local indexes of cancer survival

National index

Stata code (version 15) implementing the modelling strategy defined in section [4.6.1](#) for the estimation of the index of cancer survival for England.

```
* ANALYSIS PROGRAM: Excess hazard modelling strategy

* Fit models for each cancer, for men and women separately
foreach cancer in "lung" "pancreas" "hodgkin" "NHL" "bladder" "brain"
"breast" "cervix" "colon" "kidney" "larynx" "leukaemia" "melanoma" "myeloma"
"oesophagus" "others" "ovary" "prostate" "rectum" "stomach" "testis" "uterus" {
    use "\\cancer'.dta", clear
    di "cancer='cancer'"

* Merge data a priori with life tables
    gen age=int(ageout)
    replace age=99 if ageout>99
    sort age sex _year dep gor country
    merge m:1 age sex _year dep gor country using "Life_table"
    drop if _merge==2
    assert _merge!=1
    drop _merge
```

```

* Set survival time data
stset finmdy, failure(dead) origin(time diagmdy) exit(time finmdy)

* Loop for analysis for men and women
qui sum sex
local sexmin=r(min)
local sexmax=r(max)
local i 'sexmin'
if 'sexmin'=='sexmax' {local s='sexmax'}
if 'sexmin'<'sexmax' {local s='sexmax'+1}
while 'i'<='s' & 'i'!= 3 {display "sex="'i'
di "Cancer is " 'cancer' " and sex 'i'"
preserve
keep if sex=='i'
count

* Generate splines by cancer type for the continuous variable age at diagnosis
* year of diagnosis, and the interaction between these two variables
if (" 'cancer' " == "testis") {
rcsgen ageddiag, knots(15 35 99) gen(rcs_age) orthog}

if (" 'cancer' " == "leukaemia") {
rcsgen ageddiag, knots(15 45 75 99) gen(rcs_age) orthog}

if (" 'cancer' " == "hodgkin") {
rcsgen ageddiag, knots(15 25 65 99) gen(rcs_age) orthog}

if (" 'cancer' " == "cervix" | " 'cancer' " == "melanoma") {
rcsgen ageddiag, knots(15 35 65 99) gen(rcs_age) orthog}

if (" 'cancer' " == "brain" | " 'cancer' " == "ovary") {
rcsgen ageddiag, knots(15 40 65 99) gen(rcs_age) orthog}

if (" 'cancer' " == "NHL" | " 'cancer' " == "breast" | " 'cancer' " == "colon" |
" 'cancer' " == "uterus") {
rcsgen ageddiag, knots(15 50 70 99) gen(rcs_age) orthog}

if (" 'cancer' " == "bladder" | " 'cancer' " == "kidney" | " 'cancer' " == "larynx" |
" 'cancer' " == "myeloma" | " 'cancer' " == "oesophagus" | " 'cancer' " == "others" |
" 'cancer' " == "prostate" | " 'cancer' " == "rectum" | " 'cancer' " == "stomach") {
rcsgen ageddiag, knots(15 65 99) gen(rcs_age) orthog}

```

```

if ("‘cancer’" == "lung" | "‘cancer’" == "pancreas") {
  rcsgen ageddiag, knots(15 40 65 75 99) gen(rcs_age) orthog}

rcsgen ydiag, df(3) gen(rcs_ydiag) orthog
gen inter_age_ydiag=ageddiag*ydiag
rcsgen inter_age_ydiag, df(3) gen(rcs_intageydiag) orthog

if ("‘cancer’" == "pancreas") {local df="2"}
else if ("‘cancer’" == "lung") {local df="3" }
else { local df="4"}
di ‘df’
estimates drop _all

* Defining candidate models

* Model 1. with non-linear and non-proportional effects of age
* and year of diagnosis, and a non linear and non proportional
* interaction between age and year of diagnosis

cap stpm2 rcs_age* rcs_ydiag* rcs_intageydiag*, scale(hazard)
bhazard(rate) df(‘df’) tvc(rcs_age* rcs_ydiag* rcs_intageydiag*)
dftvc(3) iterate(20)
local error1=_rc
local ConvergedModel1=e(converged)
if ‘ConvergedModel1’==1 & ‘error1’==0{
  local Model1AIC=e(AIC)
  estimates store Model1_ENGLAND_‘cancer’_‘i’ }
else local Model1AIC=.

* Model 2. with non-linear and non-proportional effects of age
* and year of diagnosis, and a non linear interaction between
* age and year of diagnosis

cap stpm2 rcs_age* rcs_ydiag* rcs_intageydiag*, scale(hazard)
bhazard(rate) df(‘df’) tvc(rcs_age* rcs_ydiag*) dftvc(3)
iterate(20)
local error2=_rc
local ConvergedModel2=e(converged)
if ‘ConvergedModel2’==1 & ‘error2’==0{
  local Model2AIC=e(AIC)
  estimates store Model2_ENGLAND_‘cancer’_‘i’}
else local Model2AIC=.

```



```
* Model 3. with non-linear and non-proportional effects of age
* and year of diagnosis
```

```
cap stpm2 rcs_age* rcs_ydiag*, scale(hazard) bhazard(rate)
df('df') tvc(rcs_age* rcs_ydiag*) dftvc(3) iterate(20)
local error3=_rc
local ConvergedModel3=e(converged)
if 'ConvergedModel3'==1 & 'error3'==0{
local Model3AIC=e(AIC)
estimates store Model3_ENGLAND_ 'cancer'_'i'}
else local Model3AIC=.
```

```
* Model 4. with non-linear and non-proportional effects of age and
* non linear year of diagnosis
```

```
cap stpm2 rcs_age* rcs_ydiag*, scale(hazard) bhazard(rate)
df('df') tvc(rcs_age*) dftvc(3) iterate(20)
local error4=_rc
local ConvergedModel4=e(converged)
if 'ConvergedModel4'==1 & 'error4'==0{
local Model4AIC=e(AIC)
estimates store Model4_ENGLAND_ 'cancer'_'i'}
else local Model4AIC=.
```

```
* Model 5. with non-linear and non-proportional effect of age and
* linear and proportional year of diagnosis
```

```
cap stpm2 rcs_age* ydiag, scale(hazard) bhazard(rate) df('df')
tvc(rcs_age*) dftvc(3) iterate(20)
local error5=_rc
local ConvergedModel5=e(converged)
if 'ConvergedModel5'==1 & 'error5'==0{
local Model5AIC=e(AIC)
estimates store Model5_ENGLAND_ 'cancer'_'i' }
else local Model5AIC=.
```

```
* Model 6. with linear effect of age and year of diagnosis and
* non-proportional effect of age
```

```
cap stpm2 ageddiag ydiag, scale(hazard) tvc(ageddiag) bhazard(rate)
df('df') dftvc(3) iterate(20)
local error6=_rc
local ConvergedModel6=e(converged)
```

```

if 'ConvergedModel6'==1 & 'error6'==0{
  local Model6AIC=e(AIC)
  estimates store Model6_ENGLAND_ 'cancer'_'i'}
else local Model6AIC=.

* Model 7. with non-linear effect of age and year of diagnosis,
* and a non-proportional effect of year of diagnosis and a non-proportional
* interaction between age and year of diagnosis

cap stpm2 rcs_age* rcs_ydiag*, scale(hazard)
bhazard(rate) df('df') tvc(rcs_ydiag* inter_age_ydiag) dftvc(3) iterate(20)
local error7=_rc
local ConvergedModel7=e(converged)
if 'ConvergedModel7'==1 & 'error7'==0{
  local Model7AIC=e(AIC)
  estimates store Model7_ENGLAND_ 'cancer'_'i' }
else local Model7AIC=.

* Selecting the simplest model from the models with the smallest AIC

if ('ConvergedModel1'==1 & 'error1'==0) |
('ConvergedModel2'==1 & 'error2'==0) |
('ConvergedModel3'==1 & 'error3'==0) |
('ConvergedModel4'==1 & 'error4'==0) |
('ConvergedModel5'==1 & 'error5'==0) |
('ConvergedModel6'==1 & 'error6'==0) |
('ConvergedModel7'==1 & 'error7'==0) {

  estimates stats Model*
  local minAIC=min('Model1AIC','Model2AIC','Model3AIC','Model4AIC',
    'Model5AIC','Model6AIC','Model7AIC')
  di "Minimum AIC: 'minAIC'"
  cap matrix drop AICmaxmin
  forvalues k=1/7{
    if 'Model'k'AIC'<='minAIC' & 'error'k'==0 & 'ConvergedModel'k'==1
    {matrix AICmaxmin = (nullmat(AICmaxmin),'k')}
    di "Candidate models"
    matrix list AICmaxmin
    local AIC=max(AICmaxmin[1,1],AICmaxmin[1,2],AICmaxmin[1,3],
    AICmaxmin[1,4],AICmaxmin[1,5],AICmaxmin[1,6],AICmaxmin[1,7])
    di "Chosen model: 'AIC'"}
  estimates restore Model'AIC'_ENGLAND_ 'cancer'_'i'

```

* Prediction of net survival by age group and period of diagnosis based on the previously selected model

```
quietly {
  fillin cancer agecat period sex
  bys agecat period: gen tt=_n
  replace tt=. if tt>10
  tab tt
  foreach h in 1 2 3 4 5 6 {
    forvalues k = 0/4 {
      cap predictnl ns_age`k'`h'=predict(meansurv timevar(tt))
      if agecat==`k' & period==`h', se(ns_se`k'`h') ci
      (ns_lci_age`k'`h' ns_uci_age`k'`h'))}}
      cap egen ns=rsum(ns_age*)
      cap egen ns_lci=rsum(ns_lci_age*)
      cap egen ns_uci=rsum(ns_uci_age*)
      cap egen ns_se=rmin(ns_se*)}
      cap append using "\\E_index_stpm2_results`cancer'_NS.dta"
      save "\\E_index_stpm2_results`cancer'_NS.dta", replace
      clear all
      restore
      local i=`i'+1}}
}
```

* Constructing the index by combining all components of the index

```
tab agecat,m nol
recode agecat (0=15) (1=45) (2=55) (3=65) (4=75), gen(agecat1)
sort cancer sex agecat
merge m:1 cancer sex agecat using "\\cancer_age_sex_weights.dta"
assert _merge==3
drop _merge
```

* Standardise by age, sex and cancer

```
gen weighedNS=ns*stand_weights
bysort period time: egen NSstand=total(weighedNS)
```

* Calculate the variance, standard error and precision for the index

```
gen varNSstand=(stand_weights^2)*(ns_se^2)
bysort period time: egen varASNS=total(varNSstand)
gen seASNS=(varASNS^(1/2))
gen prec=1/varASNS
```

Local index

Stata code (version 15) implementing the modelling strategy defined in section 4.6.1 for the estimation of the index of cancer survival for each CCG.

```
* ANALYSIS PROGRAM: Excess hazard modelling strategy

* Fit models for each cancer, for men and women, and CCG separately
foreach cancer in "breast" "colorectum" "lung" "others" {
    use "\\cancer'exportccg.dta", clear
    di "cancer='cancer'"

* Merge data a priori with life tables
    gen age=int(ageout)
    replace age=99 if ageout>99
    sort age sex _year dep gor
    merge m:1 age sex _year dep gor using "Life_table"
    assert _merge!=1
    drop if _merge==2
    drop _merge

* Set survival time data
    stset finmdy, failure(dead) origin(time diagmdy2) exit(time censormdy)

* Loop for analysis for men and women
    qui sum sex
    local sexmin=r(min)
    local sexmax=r(max)
    local i 'sexmin'
    if 'sexmin'=='sexmax' {local s='sexmax'}
    if 'sexmin'<'sexmax' {local s='sexmax'+1}
    while 'i'<='s' & 'i'!= 3 {display "sex='"i'"

* Loop for CCGs
    foreach CCG in 001 002 003 004 005 006 007 008 009 010 011 012 013
    014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030
    031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047
    048 049 050 051 052 053 054 055 056 057 058 059 060 061 062 063 064
    065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080 081
    082 083 084 085 086 087 088 089 090 091 092 093 094 095 096 097 098
    099 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115
    116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132
    133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149
```

```

150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166
167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183
184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
201 202 203 204 205 206 207 208 209 210 211 {

di "Cancer is "‘cancer’", CCG is "‘CCG’" and sex ‘i’"
preserve
keep if sex==‘i’
keep if CCG=="‘CCG’"
count

* Generate splines by cancder type for the continuous variable age at diagnosis
* year of diagnosis, and the interaction between these two variables

if ("‘cancer’" == "breast" | "‘cancer’" == "colorectum") {
rcsgen ageddiag, knots(15 50 70 99) gen(rcs_age) orthog}

if ("‘cancer’" == "lung" | "‘cancer’" == "others") {
rcsgen ageddiag, knots(15 65 99) gen(rcs_age) orthog }

rcsgen ydiag, df(2) gen(rcs_ydiag) orthog
gen inter_age_ydiag=ageddiag*ydiag
rcsgen inter_age_ydiag, df(3) gen(rcs_intageydiag) orthog
estimates drop _all

* Defining candidate models

* Model 1. with non-linear and non-proportional effects of age
* and year of diagnosis, and a non linear and non proportional
* interaction between age and year of diagnosis

cap stpm2 rcs_age* rcs_ydiag* rcs_intageydiag*, scale(hazard)
bhazard(rate) df(4) tvc(rcs_age* rcs_ydiag* rcs_intageydiag*)
dftvc(3) iterate(15)
local error1=_rc
local ConvergedModel1=e(converged)
if ‘ConvergedModel1’==1 & ‘error1’==0{
local Model1AIC=e(AIC)
estimates store Model1_‘CCG’_‘cancer’_‘i’}
else local Model1AIC=.

* Model 2. with non-linear and non-proportional effects of age
* and year of diagnosis, and a non linear interaction between

```

```

* age and year of diagnosis

    cap stpm2 rcs_age* rcs_ydiag* rcs_intageydiag*, scale(hazard)
    bhazard(rate) df(4) tvc(rcs_age* rcs_ydiag*) dftvc(3)
    iterate(15)
    local error2=_rc
    local ConvergedModel2=e(converged)
    if 'ConvergedModel2'==1 & 'error2'==0{
    local Model2AIC=e(AIC)
    estimates store Model2_ 'CCG'_ 'cancer'_ 'i'}
    else local Model2AIC=.

* Model 3. with non-linear and non-proportional effects of age
* and year of diagnosis

    cap stpm2 rcs_age* rcs_ydiag*, scale(hazard) bhazard(rate)
    df(4) tvc(rcs_age* rcs_ydiag*) dftvc(3) iterate(15)
    local error3=_rc
    local ConvergedModel3=e(converged)
    if 'ConvergedModel3'==1 & 'error3'==0{
    local Model3AIC=e(AIC)
    estimates store Model3_ 'CCG'_ 'cancer'_ 'i'}
    else local Model3AIC=.

* Model 4. with non-linear and non-proportional effects of age and
* non linear year of diagnosis

    cap stpm2 rcs_age* rcs_ydiag*, scale(hazard) bhazard(rate)
    df(4) tvc(rcs_age*) dftvc(3) iterate(15)
    local error4=_rc
    local ConvergedModel4=e(converged)
    if 'ConvergedModel4'==1 & 'error4'==0{
    local Model4AIC=e(AIC)
    estimates store Model4_ 'CCG'_ 'cancer'_ 'i'}
    else local Model4AIC=.

* Model 5. with non-linear and non-proportional effect of age and
* linear and proportional year of diagnosis

    cap stpm2 rcs_age* ydiag, scale(hazard) bhazard(rate) df(4)
    tvc(rcs_age*) dftvc(3) iterate(15)
    local error5=_rc
    local ConvergedModel5=e(converged)

```

```

    if 'ConvergedModel5'==1 & 'error5'==0{
        local Model5AIC=e(AIC)
        estimates store Model5_ 'CCG'_ 'cancer'_ 'i' }
    else local Model5AIC=.

* Model 6. with linear effect of age and year of diagnosis and
* non-proportional effect of age

    cap stpm2 ageddiag ydiag, scale(hazard) tvc(ageddiag) bhazard(rate)
    df(4) dftvc(3) iterate(15)
    local error6=_rc
    local ConvergedModel6=e(converged)
    if 'ConvergedModel6'==1 & 'error6'==0{
        local Model6AIC=e(AIC)
        estimates store Model6_ 'CCG'_ 'cancer'_ 'i' }
    else local Model6AIC=.

* Model 7. with non-linear effect of age and year of diagnosis,
* and a non-proportional effect of year of diagnosis and a non-proportional
* interaction between age and year of diagnosis

    cap stpm2 rcs_age* rcs_ydiag*, scale(hazard)
    bhazard(rate) df('df') tvc(rcs_ydiag* inter_age_ydiag) dftvc(3) iterate(20)
    local error7=_rc
    local ConvergedModel7=e(converged)
    if 'ConvergedModel7'==1 & 'error7'==0{
        local Model7AIC=e(AIC)
        estimates store Model7_ 'CCG'_ 'cancer'_ 'i' }
    else local Model7AIC=.

* Model 8. Model with non-linear effect of age and year of
* diagnosis and a non-linear and non-proportional interaction
* between age and year of diagnosis

    cap stpm2 rcs_age* rcs_ydiag* inter_age_ydiag, scale(hazard)
    bhazard(rate) df('df') tvc(rcs_intageydiag*) dftvc(3) iterate(20)
    local error8=_rc
    local ConvergedModel8=e(converged)
    if 'ConvergedModel8'==1 & 'error8'==0{
        local Model8AIC=e(AIC)
        estimates store Model8_ 'CCG'_ 'cancer'_ 'i' }
    else local Model8AIC=.

```

```

* Selecting the simplest model from the models with the smallest AIC

if ('ConvergedModel1'==1 & 'error1'==0) |
('ConvergedModel2'==1 & 'error2'==0) |
('ConvergedModel3'==1 & 'error3'==0) |
('ConvergedModel4'==1 & 'error4'==0) |
('ConvergedModel5'==1 & 'error5'==0) |
('ConvergedModel6'==1 & 'error6'==0) |
('ConvergedModel7'==1 & 'error7'==0) |
('ConvergedModel8'==1 & 'error8'==0) {

    estimates stats Model*
    local minAIC=min('Model1AIC','Model2AIC','Model3AIC','Model4AIC',
    'Model5AIC','Model6AIC','Model7AIC','Model8AIC')
    di "Minimum AIC: 'minAIC'"
    cap matrix drop AICmaxmin
    forvalues k=1/8{
    if 'Model'k'AIC'<='minAIC' & 'error'k'==0 & 'ConvergedModel'k'==1{

        matrix AICmaxmin = (nullmat(AICmaxmin),'k')
        di "Candidate models"

        matrix list AICmaxmin
        local AIC=max(AICmaxmin[1,1],AICmaxmin[1,2],AICmaxmin[1,3],
        AICmaxmin[1,4],AICmaxmin[1,5],AICmaxmin[1,6],AICmaxmin[1,7],
        AICmaxmin[1,8])
        di "Chosen model: 'AIC'"}}
        estimates restore Model'AIC'_'CCG'_'cancer'_'i'

* Prediction of net survival by age group and year of diagnosis based
on the previously selected model

quietly {
    fillin cancer agecat ydiag sex CCG
    bys agecat ydiag: gen tt=_n
    replace tt=. if tt>5
    tab tt
    foreach h in 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
    2007 2008 2009 2010 2011 {
    forvalues k = 0/4 {
    cap predictnl ns_age'k' 'h'=predict(meansurv timevar(tt)) if agecat=='k' ///

    & ydiag=='h', se(ns_se'k' 'h') ci (ns_lci_age'k' 'h' ns_uci_age'k' 'h')

```



```

    }}}
    cap egen ns=rsum(ns_age*)
    cap egen ns_lci=rsum(ns_lci_age*)
    cap egen ns_uci=rsum(ns_uci_age*)
    cap egen ns_se=rmin(ns_se*)
  }
  cap append using "\\CCG_stpm2_results'cancer'.dta"
  save "\\CCG_stpm2_results'cancer'.dta", replace
  clear all
  restore
}
local i='i'+1 }}

* Constructing the index by combining all components of the index
tab agecat,m nol
recode agecat (0=15) (1=45) (2=55) (3=65) (4=75), gen(agecat1)
sort cancer sex agecat

merge m:1 cancer sex agecat using "\\cancer_age_sex_weights.dta"
assert _merge==3
drop _merge

* Standardise by age, sex and cancer
gen weighedNSCCG=ns*stand_weights
bysort CCG ydiag: egen NSstandCCG=total(weighedNSCCG)

* Calculate the variance, standard error and precision for the index
gen varNSstandCCG=(stand_weights^2)*(ns_se^2)
bysort CCG ydiag: egen varASNSCCG=total(varNSstandCCG)
gen seASNSCCG=(varASNSCCG^(1/2))
gen prec=1/varASNSCCG

```

A.2 R code to construct a funnel plot

R code example to construct a funnel plot using the estimates of the index of cancer survival for CCGs.

```
* Load data file containing estimates for the index of cancer survival
  setwd("../funnel plots")
  getwd()

  data_CCG_all_cancers_all_ages_temp<-read.table("../Data_funnel_plots.txt",
  header=TRUE,sep="\t")

  data_CCG_all_cancers_all_ages<-data.frame(data_CCG_all_cancers_all_ages_temp)

* Construct funnel plot
  plot(data_CCG_all_cancers_all_ages$prec,data_CCG_all_cancers_all_ages$ns,
  frame = FALSE,font.lab=2,pch=21,bg="black",axes=FALSE,xlab="Precision",
  ylab="One-year index of net survival(%)",
  xlim=c(0,20),ylim=c(50,80),cex.main=1.5)
  axis(2, c(50,55,60,65,70,75,80),las=1,font.axis=2,tck=-0.03)
  axis(1, c(0,5,10,15,20),las=1,font.axis=2,tck=-0.03)

* Define and draw the target
  NSE<-data_CCG_all_cancers_all_ages$nsE[1]
  x<-seq(0,20,by=1)
  baseline<-rep(NSE,21)
  lines(x,baseline,lwd=2,col="black")

* Define the control limits
  denom<-function(x,y,z){
    seq(x,y,by=z)}

* Lower 95% Control Limit

* Precision
  denom_ll95_temp<-denom(0,20,0.1)

* Variance
  denom_ll95<-(denom_ll95_temp)^(-1)

* SE
  denom_ll95SE<-sqrt((denom_ll95_temp)^(-1))
  llimit_95<-exp(-exp(log(-log(NSE/100))+1.96*sqrt((((denom_ll95SE/100)^2))/
  ((NSE/100*log (NSE/100))^2))))*100
  lines(denom_ll95_temp,llimit_95,lty=2, col="grey")
```

```

* Upper 95% Control Limit

* Precision
denom_ul95_temp<-denom(0,20,0.1)
* Variance
denom_ul95<-(denom_ul95_temp)^(-1)
* SE
denom_ul95SE<-sqrt((denom_ul95_temp)^(-1))
ulimit_95<-exp(-exp(log(-log(NSE/100))-1.96*sqrt((((denom_ul95SE/100)^2))/
((NSE/100*log (NSE/100))^2))))*100
lines(denom_ul95_temp,ulimit_95,lty=2, col="grey")

* Lower 99,8% Control Limit

* Precision
denom_ll998_temp<-denom(0,20,0.1)
* Variance
denom_ll998<-(denom_ll998_temp)^(-1)
* SE
denom_ll998SE<-sqrt((denom_ll998_temp)^(-1))
llimit_998<-exp(-exp(log(-log(NSE/100))+3.09*sqrt((((denom_ll998SE/100)^2))/
((NSE/100*log (NSE/100))^2))))*100
lines(denom_ll998_temp,llimit_998,lty=3)

* Upper 99,8% Control Limit

* Precision
denom_ul998_temp<-denom(0,20,0.1)
* Variance
denom_ul998<-(denom_ul998_temp)^(-1)
* SE
denom_ul998SE<-sqrt((denom_ul998_temp)^(-1))
ulimit_998<-exp(-exp(log(-log(NSE/100))-3.09*sqrt((((denom_ul998SE/100)^2))/
((NSE/100*log (NSE/100))^2))))*100
lines(denom_ul998_temp,ulimit_998,lty=3)

* Control limits legend
legend(10, 54, "95% control limits", lty=2, cex=1.2 ,bty = "n")
legend(10, 52, "99.8% control limits", lty=3, cex=1.2 ,bty = "n")

```

```
* Code to identify points below the lower 99,8% control limit

llimit_998detectoutliers<-exp(-exp(log(-log(NSE/100))+3.09*
sqrt((1/(data_CCG_all_cancers_all_ages$prec*(100^2)))/
((NSE/100*log (NSE/100))^2))))*100

flatllimitoutlier<-0

data_outliers<-data.frame(data_CCG_all_cancers_all_ages,llimit_998detectoutliers,
flatllimitoutlier)

for (i in 1:nrow(data_outliers)){
  if (data_outliers$ns[i]<data_outliers$llimit_998detectoutliers[i])
  {data_outliers$flatllimitoutlier[i]<-1
  points(data_outliers$prec[i],data_outliers$ns[i],pch=21,bg="darkorange3",
  cex = 1)}}}
```

A.3 R code to implement flexible Bayesian excess hazard models using low-rank thin plate splines

R code example to implement the flexible Bayesian excess hazard model proposed in research publication 3 [\[257\]](#).

```
# Log-excess hazard model using low-rank thin plate splines

# Model specification
write("model{ # Start model definition
for(i in 1:N) { # Open loop for individual observations
  tlp[i,1:(K+1)] <- alpha[1,] + alpha[2,]*agediagi.cent[i]
  eta[i] <- inprod(tlp[i,],T[i,]) + inprod(beta[],X[i,]) + epsilon*dep[i]
  # Loop over the k splittings of follow-up time to define the cumulative
  excess hazard function
  for(k in 1:K){
    Haz[i,k] <- exp(inprod(tlp[i,],TT[i,k,]))*(1-exp(-(TT[i,k,2]-tilde.t[k])*
      inprod(tlp[i,],U[k,])))/inprod(tlp[i,],U[k,])
  }
  # Negative log-likelihood function
  neg.LL[i] <- max(-delta[i]*log(brate[i] + max(exp(eta[i]), 0.0001)) +
    sum(Haz[i,])*exp(inprod(beta[],X[i,]))*exp(epsilon*dep[i]) + C, 0.0001)
    zeros[i] ~ dpois(neg.LL[i]);
  # Use Poisson distribution with zeros trick to specify non-standard likelihood
} # Close loop for individual observations

# Prior on baseline excess hazard parameters
# and hyperpriors on variance parameters
for(q in 1:2){
  alpha[q,1] ~ dnorm(0,0.0001)
  alpha[q,2] ~ dnorm(0,0.0001)
  for(k in 2:K){alpha[q,k+1] ~ dnorm(0,tau.alpha[q])}
  tau.alpha[q] <- 1/(sigma.alpha[q]*sigma.alpha[q])
  sigma.alpha[q] ~ dunif(0.01,100)
}

# Prior on regression parameters
# and hyperpriors on variance parameters
beta[1] ~ dnorm(0,0.0001)
for(j in 2:J){ beta[j] ~ dnorm(0,tau.beta)}
epsilon ~ dnorm(0,0.0001)
tau.beta <- 1/(sigma.beta*sigma.beta)
```

```

sigma.beta ~ dunif(0.01,100)
}","Model_London_colon_men.txt")

# Load data #
dataset = read.table("Colon_London_with_stage_2009_men.txt", sep="\t", header=T)
names(dataset)

# Define cut-points for follow-up time to use as spline partitions
min(t)
max(t)
K=4
tilde.t = c(0, 0.18, 0.84, 2.26, 6)

# Example to define equally spaced partition
#K = 4 # Number of splits of follow-up time interval
#tilde.t = seq(min(t),max(t),length.out=5)

# Define spline for the baseline excess hazard #

# Time Transformation Matrix - does not depend on the actual observed times but
# it is computed from the time partition vector
OMEGA_alpha <- abs(outer(tilde.t[-c(1,K+1)],tilde.t[-c(1,K+1)],"-"))
svd.OMEGA_alpha <- svd(OMEGA_alpha)
sqrt.OMEGA_alpha <- t(svd.OMEGA_alpha$v%*(t(svd.OMEGA_alpha$u)*
  sqrt(svd.OMEGA_alpha$d)))
inv.D <- solve(cbind(c(1,rep(0,K)),c(0,1,rep(0,K-1)),rbind(rep(0,K-1),rep(0,K-1),
  sqrt.OMEGA_alpha)))

# Construct Time Design Matrices - for observed time
T_K <- cbind(1,t(sapply(t, function(x) abs(x - tilde.t[-c(K+1)]) -
  abs(tilde.t[-c(K+1)]))))
T = T_K%*%inv.D

# Construct Time Design Matrices for the cumulative excess hazard
tk = t(sapply(t,function(z) pmax(pmin(z,tilde.t[-1]),tilde.t[-(K+1)])))
TT_K = TT = array(NA,c(N,K,K+1))

# TT_K is equivalent to T_K but for each time partition
# TT is equivalent to T but for each time partition
for(k in 1:K) TT_K[,k,] <- cbind(1,t(sapply(tk[,k], function(z)
  abs(z - tilde.t[-c(K+1)]) - tilde.t[-c(K+1)])))
for(i in 1:N) TT[i,,] = TT_K[i,,]%*%inv.D

```

```

U_K <- matrix(1, K, K)
U_K[upper.tri(U_K)] <- -1
U <- U_K%*%inv.D[-1,]

# Define splines on covariates
# Covariate Range Partition for age at diagnosis centered and reduced
J=3
tilde.agediag = seq(min(dataset$agediag.cent.red),max(dataset$agediag.cent.red),
  length.out=J+1)

# Transformation/Penalty Matrix
OMEGA_beta<-rbind(c(1,rep(0,J-1)),cbind(0,abs(outer(tilde.agediag[-c(1,J+1)],
  tilde.agediag[-c(1,J+1)],"-"))^3))
svd_OMEGA_beta<-svd(OMEGA_beta)
inv.D_beta<-solve(t(svd_OMEGA_beta$v%*(t(svd_OMEGA_beta$u)*
  sqrt(svd_OMEGA_beta$d))))

# Design Matrix
X = cbind(dataset$agediag.cent.red,t(sapply(dataset$agediag.cent.red,function(z)
  abs(z - tilde.agediag[-c(1,J+1)])^3-abs(tilde.agediag[-c(1,J+1)])^3))))%*%
  inv.D_beta

# Define data (dta), initial values (ints) and parameters
# to monitor (pars) in Jags

dta <- list("N", "K", "J", "tilde.t", "T", "TT", "U", "delta", "X", "zeros",
  "C", "brate", "agediagi.cent", "dep")
pars <- c("alpha", "tau.alpha", "sigma.alpha", "tau.beta", "beta",
  "sigma.beta", "epsilon")

inits1 <- list(sigma.alpha=runif(2,0.01,100), alpha=matrix(rnorm(2*(K+1)),2,K+1),
  sigma.beta=runif(1,0.01,1), beta=c(0,0,0), epsilon=0)
inits2 <- list(sigma.alpha=runif(2,0.01,100), alpha=matrix(rnorm(2*(K+1)),2,K+1),
  sigma.beta=runif(1,0.01,1), beta=c(0.01,0.01,0.01), epsilon=1)
ints <- list(inits1, inits2)

# Call JAGS using R2Jags
Sys.time() # Set time monitor
NetSurv.fit <- jags(data=dta,inits=ints,model.file="Model_London_colon_men.txt",
  parameters=pars, n.chains=2,n.iter=50000,n.burnin=5000,n.thin=3)
Sys.time()

```

```

# print summary of posterior distribution for parameters and traceplots
print(NetSurv.fit)
traceplot(NetSurv.fit)

# Extract parameter chains for post-prediction:
NetSurv.fit.chains <- as.mcmc(NetSurv.fit)
summary(NetSurv.fit.chains[1])
summary(NetSurv.fit.chains[2])

# Create a matrix with the chains and appends the 2 chains
chains.matrix1 <- as.matrix(NetSurv.fit.chains[1])
dim(chains.matrix1)
summary(chains.matrix1)
chains.matrix2 <- as.matrix(NetSurv.fit.chains[2])
dim(chains.matrix2)
summary(chains.matrix2)

# Append results of the two chains
chains.matrix <- rbind(chains.matrix1, chains.matrix2)
dim(chains.matrix)
head(chains.matrix)

# reorder the columns of the matrix:
list <- c("alpha[1,1]", "alpha[1,2]", "alpha[1,3]", "alpha[1,4]", "alpha[1,5]",
  "alpha[2,1]", "alpha[2,2]", "alpha[2,3]", "alpha[2,4]", "alpha[2,5]", "beta[1]",
  "beta[2]", "beta[3]", "epsilon", "sigma.alpha[1]", "sigma.alpha[2]",
  "tau.alpha[1]", "tau.alpha[2]", "sigma.beta", "tau.beta", "deviance")
chains.matrix.reordered <- chains.matrix[,list]
head(chains.matrix.reordered)

# Saving the chains for future use:
write.table(chains.matrix.reordered, file="Chains_Netsurv.csv")

# Use saved chains
chains.matrix.reordered1 = read.csv("Chains_Netsurv.csv", header=TRUE, sep="")
head(chains.matrix.reordered1)
chains.matrix.reordered <- as.matrix(chains.matrix.reordered1)
is.matrix(chains.matrix.reordered)
dim(chains.matrix.reordered)
colnames(chains.matrix.reordered) <- c("alpha[1,1]", "alpha[1,2]", "alpha[1,3]",
  "alpha[1,4]", "alpha[1,5]", "alpha[2,1]", "alpha[2,2]", "alpha[2,3]", "alpha[2,4]",
  "alpha[2,5]", "beta[1]", "beta[2]", "beta[3]", "epsilon", "sigma.alpha[1]",
  "sigma.alpha[2]", "tau.alpha[1]", "tau.alpha[2]", "sigma.beta", "tau.beta", "deviance")

```



```
# Save chains for each parameter
chains.matrix.par.temp <- chains.matrix.reordered[, -c(K+K+J+2+2) : -c(K+K+J+2+2+6)]
chains.matrix.par <- chains.matrix.par.temp
is.matrix(chains.matrix.par)
head(chains.matrix.par)
dim(chains.matrix.par)

# Alpha parameters
chains.matrix.par.alpha <- chains.matrix.par[, c(-(K+K+2+1) : -(K+K+2+1+J+2))]
is.matrix(chains.matrix.par.alpha)
head(chains.matrix.par.alpha)
dim(chains.matrix.par.alpha)

chains.matrix.par.alpha_1 <- chains.matrix.par.alpha[, c(-(K+2) : -(K+6))]
dim(chains.matrix.par.alpha_1)
head(chains.matrix.par.alpha_1)

chains.matrix.par.alpha_2 <- chains.matrix.par.alpha[, c(-(1) : -(K+1))]
dim(chains.matrix.par.alpha_2)
head(chains.matrix.par.alpha_2)

# Beta parameters
chains.matrix.par.beta.temp <- chains.matrix.par[, c(-(1) : -(K+K+1+1))]
chains.matrix.par.beta <- chains.matrix.par.beta.temp[, c(-(J+1) : -(J+1))]
is.matrix(chains.matrix.par.beta)
head(chains.matrix.par.beta)
dim(chains.matrix.par.beta)

# Epsilon parameter
chains.matrix.par.epsilon <- as.matrix(chains.matrix.par[, c(-(1) : -(K+K+2+J))])
is.matrix(chains.matrix.par.epsilon)
head(chains.matrix.par.epsilon)
dim(chains.matrix.par.epsilon)
```

```

# Post-estimation function to estimate net survival

# Create prediction time
predtime <- seq(0.1,5.99,0.1)
L=length(predtime)

# Number of effective sampled values to be used
samples=30000

# Creates a sequence for the observed age range which for this dataset is 16-99
predage <- seq(16, 99, 1)
A=length(predage)

# Center and reduce the age prediction vector
predage.cent <- (predage - 70)/100

# Creates a matrix to expand the prediction age vector for each deprivation
category (dimension AD=A*5=420)

predep <- c(1,2,3,4,5)
predagedep <- expand.grid(age=predage.cent, dep=predep)
dim(predagedep)
AD=length(predagedep[,1])

# Creates empty matrix with dimension AD*(no. time points L) with NA's:
pred_matrix <- matrix(data=NA, nrow=AD, ncol=L)
dim(pred_matrix)

# Fill in prediction matrix with the prediction times:
for(i in 1:L) pred_matrix[,i] <- rep(predtime[i],times=AD)

# Construct the time design matrix for the prediction
OMEGA_alpha <- abs(outer(tilde.t[-c(1,K+1)],tilde.t[-c(1,K+1)],"-"))
svd.OMEGA_alpha <- svd(OMEGA_alpha)
sqrt.OMEGA_alpha <- t(svd.OMEGA_alpha$v%*(t(svd.OMEGA_alpha$u)*
  sqrt(svd.OMEGA_alpha$d)))
inv.D <- solve(cbind(c(1,rep(0,K)),c(0,1,rep(0,K-1)),rbind(rep(0,K-1),
  rep(0,K-1),sqrt.OMEGA_alpha)))

# Create a 3 dimensional array to include the several time prediction points L
T_K_pred = array(NA, c(AD, K+1, L))
for(l in 1:L) T_K_pred[, ,l] <- pred_matrix[,l]
for(l in 1:L) T_K_pred[, ,l] = cbind(1,t(sapply(pred_matrix[,l],function(z)
  abs(z - tilde.t[-c(K+1)]) - tilde.t[-c(K+1)])))

```

```

# T_K_pred had dimension (AD,k+1,L)
# inv.D has dimension (k+1,k+1)
# T_pred has dimension (AD, k+1, L)

T_pred = array(NA, c(AD, K+1, L))
for(l in 1:L) T_pred[, ,l] = T_K_pred[, ,l]%%inv.D

# Construct the time design matrices the cumulative hazard for the prediction

# tk_pred has dimension (AD,K,L)
# create a 3 dimensional array to include the several time prediction points L
tk_pred = array(NA, c(AD, K, L))

for(l in 1:L) tk_pred[, ,l] <- pred_matrix[,l]
for(l in 1:L) tk_pred[, ,l] = t(sapply(pred_matrix[,l],function(z)
  pmax(pmin(z,tilde.t[-1]),tilde.t[-(K+1)])))

# creates a 4 dimensional array
TT_K_pred = TT_pred = array(NA,c(length(pred_matrix[,1]),K,K+1,L))
for(l in 1:L) TT_K_pred[, , ,l] <- pred_matrix[,l]
for(l in 1:L) TT_pred[, , ,l] <- pred_matrix[,l]
for(l in 1:L){for(k in 1:K) TT_K_pred[,k, ,l] = cbind(1,t(sapply(tk_pred[,k,l],
  function(z) abs(z - tilde.t[-c(K+1)]) - tilde.t[-c(K+1)])))}

# TT_K_pred has dimension (AD, k, k+1, L)
# inv.D has dimension (k+1,k+1)
# TT_pred has dimension (AD, K, k+1, L)

for(l in 1:L){for(i in 1:length(pred_matrix[,1])) TT_pred[i, , ,l] =
  TT_K_pred[i, , ,l]%%inv.D}

U_K = matrix(1,K,K)
U_K[upper.tri(U_K)] <- -1;
U <- U_K%%inv.D[-1,]

# Covariate Range Partition
J=3
tilde.agediag = seq(min(predage.cent),max(predage.cent),length.out=4)

```

```

# Transformation/Penalty Matrix
OMEGA_beta<-rbind(c(1,rep(0,J-1)),cbind(0,abs(outer(tilde.agediag[-c(1,J+1)],
  tilde.agediag[-c(1,J+1)],"-"))^3))
svd_OMEGA_beta<-svd(OMEGA_beta)
inv.D_beta<-solve(t(svd_OMEGA_beta$v%*(t(svd_OMEGA_beta$u)*sqrt(svd_OMEGA_beta$d))))

# Design Matrix
X = cbind(predagedep[,1],t(sapply(predagedep[,1],function(z)
  abs(z - tilde.agediag[-c(1,J+1)])^3-abs(tilde.agediag[-c(1,J+1)])^3))))%*%inv.D_beta

# Deriving the survival function for each observation

# Transpose the result of applying the function over all the sampled values
# and applying the function over all the observations [length(pred_matrix[,1])]
# dim(chains.matrix.par.alpha): (no. sampled values, no. of estimated parameters)
# dim(chains.matrix.par.alpha)[1]: no. sampled values

# Dimension Surv_ind (no. sampled values, no.observations)

Surv_ind=array(NA, c(samples, length(pred_matrix[,1]), L))
for(l in 1:L) Surv_ind[,1] <- pred_matrix[,1]
dim(Surv_ind)

Sys.time()
for(l in 1:L){
  Surv_ind[,1] = t(sapply(1:dim(chains.matrix.par.alpha)[1],
    function(r) sapply(1:length(pred_matrix[,1]),
      function(i) exp(-exp(predagedep[i,2]%*%chains.matrix.par.epsilon[r])*
        exp(X[i,]%*%chains.matrix.par.beta[r,])*
        sum(exp(TT_pred[i,,1] %*% (cbind(chains.matrix.par.alpha_1[r,]) +
          cbind(chains.matrix.par.alpha_2[r,]) %*% predagedep[i,1]))) *
        (1-exp(-(TT_pred[i,,2]-tilde.t[-(K+1)])) *
          (U%*(cbind(chains.matrix.par.alpha_1[r,]) + cbind(chains.matrix.par.alpha_2[r,])
            %*% predagedep[i,1])))))/
        (U%*(cbind(chains.matrix.par.alpha_1[r,]) + cbind(chains.matrix.par.alpha_2[r,])
          %*% predagedep[i,1])) ) ) ) )
  }
Sys.time()

```

```
# Average over observations for each sampled value from the chain to obtain
a net survival estimate for the whole cohort by applying a function to
calculate the mean for each L (time point), averaging over the observations
N (lines) in the Surv_ind matrix[samples, N]

Average_observations_by_samplevalue_netsurv=array(NA, c(samples, L))
for(l in 1:L) Average_observations_by_samplevalue_netsurv[,l] =
  apply(Surv_ind[,l], 1, function(x) mean(x))

# Calculate summary statistics for each of the L posterior distributions
of net survival by applying the mean and quantiles to the columns of
the matrix Average_observations_by_samplevalue_netsurv [samples * L]

Net_surv = array(NA, c(3, L))
dim(Net_surv)
for(l in 1:L) Net_surv = apply(Average_observations_by_samplevalue_netsurv,
  2, function(x) c(mean(x), quantile(x, c(0.025,0.975))) ) )
```

Appendix B

Other relevant research activities undertaken

B.1 Research degree student poster day

Poster submitted to the research degree student poster day and awarded the first prize for the Department of Epidemiology and Population Health (March 2013).



Small-area estimation of cancer survival

Manuela Quaresma, CRUK Cancer Survival Group, EPH, LSHTM
Supervisor: Dr. Bernard Rachet



Background

Relative survival is the main measure of cancer survival reported by population-based cancer registries. It quantifies the **excess mortality** in cancer patients after correction for other causes of death. Of special interest is the estimation of cancer **survival** at a **small area** level to:

Aim A: help guide local policy for cancer care

Aim B: be used as a national tool for surveillance

Objectives

1. Estimate and identify **patterns** of **geographical** and **temporal variation** in cancer survival at small-area level
2. Develop the application of funnel plots to explore regional and temporal variations in relative survival
3. Develop mapping techniques to visualise regional variations in relative survival

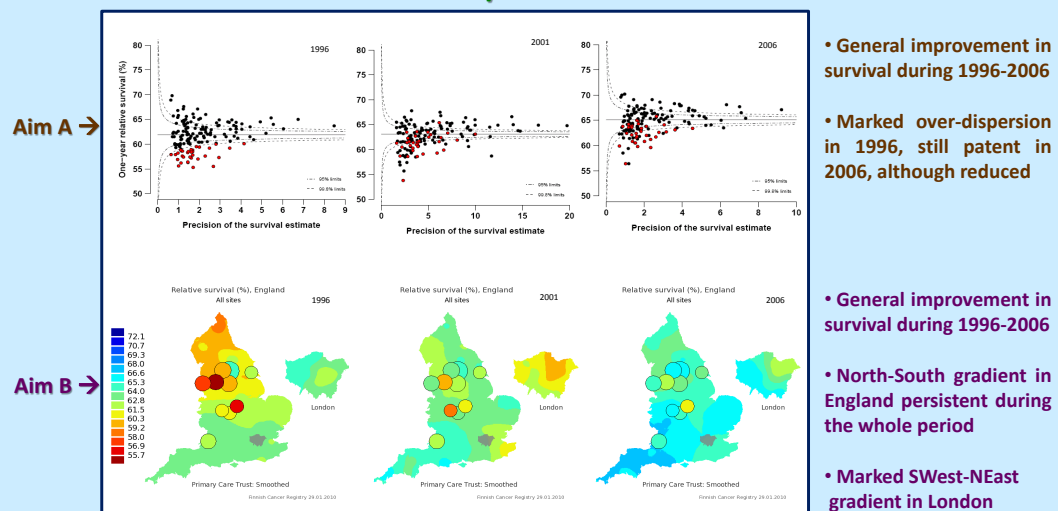
Methods

- **Flexible modelling** of the cumulative **excess hazard** (splines)
- **Funnel plots** of the individual small-area estimates → **Aim A**
- **Smoothed maps** using floating weighted averages and weights defined as the inverse of distance → **Aim B**
Large cities not smoothed (colored circles)

Application

- **152 Primary Care Trusts (PCT)** in England → sparse data
- **Data:** all adults diagnosed during 1997-2006 in England with a **cancer** and followed up until 2007
- **One-year relative survival** for all cancers **combined**
- **Adjustment** for differences in the distribution of **age, sex and type of cancer**

Preliminary results



Problem: small percentage of missing estimates in some PCT → non-convergence

Further developments

- Small-area estimation techniques: Spatial autoregressive (**SAR**), conditional autoregressive (**CAR**) models, Bayesian approaches
- Over-dispersion techniques for funnel plots and improved smoothing techniques for mapping

B.2 Beautiful data competition

Entry submitted to the 'Beautiful data' competition organised at the London School of Hygiene & Tropical Medicine and chosen as one of the two first prize winners. The prize was a Guardian Masterclass day workshop on Data Visualisation (May 2013).

Data visualisation: funnel plots and mapping for small-area cancer survival

by Manuela Quaresma

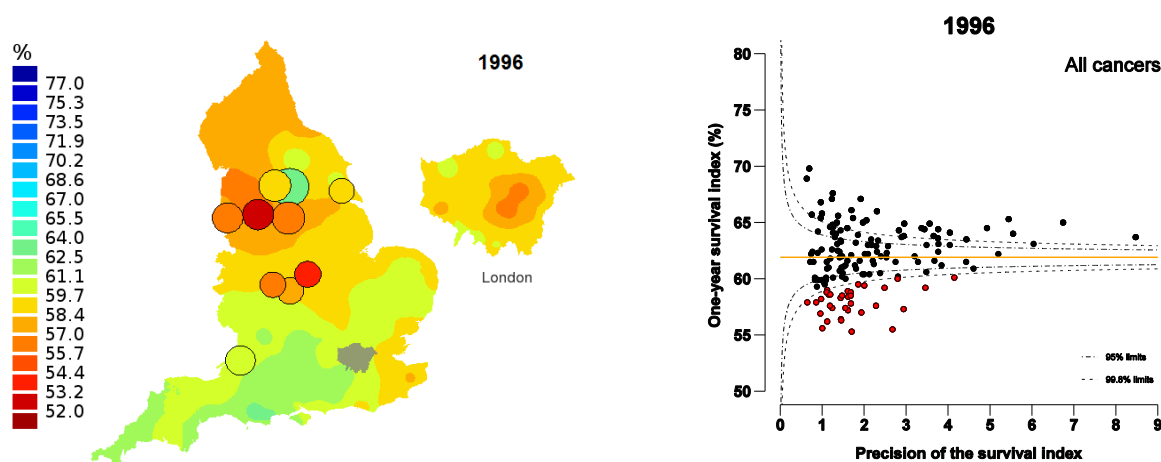
Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
+44 (0)20 7927 2856; manuela.quaresma@lshtm.ac.uk

Dissemination of cancer survival research is mainly aimed at informing patients and the public, raising awareness and influence policy, monitoring policy impacts and at prompting change. Nevertheless, dissemination of cancer survival statistics has traditionally been done using mainly tables containing listings of cancer survival statistics, typically stratified by geographical area and year of diagnosis as is exemplified in the table below:

Region	Year of diagnosis													
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
A	67	67	68	68	69	69	70	70	71	71	71	72	73	73
B	69	68	69	70	69	70	70	71	71	72	71	72	72	72
C	64	65	66	67	68	69	70	71	71	72	73	73	74	75
D	66	66	67	67	68	68	69	69	70	71	70	71	72	72
E	67	67	68	68	68	69	69	70	70	70	71	71	72	72
F	66	67	67	68	69	69	69	70	71	71	71	72	72	73
G	67	67	68	69	69	70	71	71	72	73	73	73	74	74
H	65	67	68	69	70	71	71	72	73	74	74	74	75	75

In fact, the table shown above is a simplified version of how such tables of results look in reality: most tables contain hundreds of survival estimates, and one single table can flip over many A4 pages! This form of presentation makes it very difficult (sometimes even impossible) for a non-specialist audience to understand the message we are trying to convey!

I believe that good communication of research into action relies on us researchers understanding who our audience is, what they need to know and finding the appropriate means of demonstrating our results. I am motivated to improve the dissemination of cancer survival results by finding simple and innovative ways to display such data. Below are two examples that I have implemented to communicate the same cancer survival results as displayed in the table above. I seek inspiration to continue developing new visualisation techniques for cancer survival statistics.



B.3 Oral presentations at conferences and meetings

B.3.1 North American Association of Central Cancer Registries

Selected for an oral presentation in the PhD student session at the North American Association of Central Cancer Registries (NAACCR) annual conference (June 2014).

B.3.2 All-party Parliamentary Group on Cancer annual meeting

Invited to present the results of the cancer survival index for CCGs at the annual meeting of the All-Party Parliamentary Group on Cancer (APPGC) held in Westminster (October 2014).

B.3.3 Royal statistical Society annual conference

Invited speaker in the medical statistics session 'Flexible hazard regression models for time-to-event data' at the annual conference of the Royal Statistical Society (September 2018).



Small-area cancer survival estimation: methodological challenges

Manuela Quaresma
Manuela.Quaresma@lshtm.ac.uk

Supervisor: Dr. Bernard Rachet

*Cancer Research UK Cancer Survival Group
London School of Hygiene and Tropical Medicine*

NAACCR 2014 Annual Conference

Ottawa, 25 June 2014

Small-area health geographies in England

Clinical Commissioning Groups (CCGs)



- National Health Service (NHS) organisations, created in April 2013
- Clinically lead groups of all General Practices (GPs) in a geographical area
- Responsible health care provision and policy
- 211 CCGs – mean population of 250,000

Background

- Increasing interest understanding geographic disparities in cancer survival at smaller area level
- Based on health geographies responsible for managing and delivering health care in their resident populations
- Inform national health-policy makers and local healthcare managers
 - Evaluate effectiveness of cancer control programmes
 - Public health planning and surveillance

Aim: Develop an all-cancers survival index for each CCG

Clinical Commissioning Groups (CCGs)



Requirements:

- Local measure of outcome (effectiveness of cancer services)
- National tool for surveillance and health strategy
- Responsive measure: short term survival (one-year after diagnosis)
- Statistically robust (sparse data)
- Fair representation

All-cancers survival index for each CCG

To build a single all-cancers survival index for each CCG, individual year of diagnosis and follow-up time, separate estimates of survival are required for each combination of:

- Type of cancer
- Age group
- Sex

The all-cancers survival index is then build as a weighted average of these individual components (3-way standardisation)

- Survival varies widely with age, sex and with type of cancer
- Valid assessment of survival trends for all cancers combined, survival index must account changes over time in distribution of age, sex and cancer type

Components of the index:

- Construct all-cancer survival index for each CCG (211) and year of diagnosis (16 years): 3376 survival indexes
- Number of survival estimates needed in total for all 211 CCGs and 16 years of diagnosis: 118,160 survival estimates

Estimation problems:

- Sparse data at small area level and combinations breakdown \Rightarrow estimates unstable with low precision and large random variation
- In extreme situations \Rightarrow missing estimates
- Follow-up time dimension adds complexity \Rightarrow set of patients at risk of dying, at any specific time after diagnosis, is continuously being depleted with the death or censoring of patients

Survival measure used in the index: Net Survival

- Net survival quantifies the survival after taking account of death from other causes (background mortality)
- Life-tables are used to take account wide variation in background mortality by age, sex, socio-economic status and region over time

Estimation of net survival:

- Non-parametric estimators (Pohar-Perme estimator): not feasible in this analysis due to sparseness of data

- Parametric approaches: regression models excess hazard

Overall hazard of death decomposed as a sum of the two hazards:

$$\lambda_{obs}(t) = \lambda_{net}(t) + \lambda_{exp}(t)$$

- ▶ $\lambda_{obs}(t)$ - overall observed mortality hazard
- ▶ $\lambda_{net}(t)$ - hazard due to the cancer, excess hazard of death or **net hazard**
- ▶ $\lambda_{exp}(t)$ - mortality hazard due to all other causes (expected hazard)

Population-based cancer registration data and deaths:

- Individual-level data for all adults patients (aged 15-99 years)
- First, primary, invasive malignancy in England during 1996-2011
- Vital status was updated on the 31 December 2012
- Weights: proportion of cancer patients diagnosed in England and Wales during 1996-99 in each of the combinations of age group, sex and type of cancer

Components of the index:

- Geography: 211 CCGs
- Individual years of diagnosis (1996-2011): 16 years
- Type of cancer: colorectum, breast (female), lung and other cancers combined
- Age groups: 15-44, 45-54, 55-64, 65-74, 75-99
- Sex: male and female

Flexible parametric regression models for excess hazard (net survival)

Log cumulative excess hazard (Royston and Parmar; Nelson et al.):

ln(H_net(t|x_i)) = s_0(ln(t)|kn) + x_i*beta_i

- s_0(ln(t)|kn) - restricted cubic spline function to model baseline log cumulative excess hazard

Net survival function: S_net(t) = exp(-H_net(t))

Modelling strategy:

Estimate net survival for each CCG, cancer group and sex, including:

- Age and year of diagnosis as main effects and modelled as continuous variables (account potential non-linearity)
- Interactions:
 - age and year of diagnosis
 - year of diagnosis and follow-up time (account for potential non-proportionality excess hazards over follow-up time)
 - age and follow-up time

Eight candidate models set up a priori:

Model	Age		Year of diagnosis		Interaction age and year of diagnosis	
	Non-linear	Non-proportional	Non-linear	Non-proportional	Non-linear	Non-proportional
1						
2						
3						
4						
5						
6						
7						
8						

- In total: 11,816 models needed to be fitted
- Model selection procedure was automated in STATA. Models fitted using command *spm2*
- Akaike Information Criterion (AIC) used to choose model with smallest AIC: best-fitting and simplest model (less parameters)

Overall performance of the modelling strategy

- Computational intensity: 1 month
- Out of 11,816 fitted models, 9,031 converged

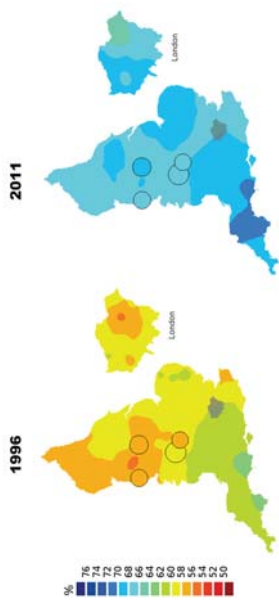
Example of patterns of model convergence (colorectum, men):

CCG		Patterns of model convergence									
		no.	%	1	2	3	4	5	6	7	8
88	42%			x							x
47	22%			x	x	x	x	x	x	x	x
29	14%			x	x	x	x	x	x	x	x
15	7%			x	x	x	x	x	x	x	x

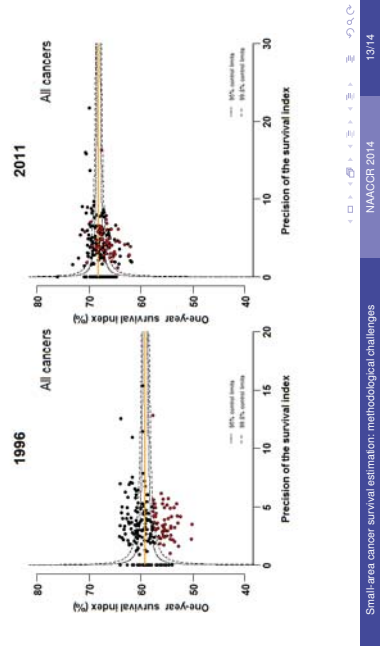
Most selected model for colorectum, 80%
Model 6: (log-) linear effect of age and year of diagnosis and non-proportional effect of age

- Missing estimates affected 5% of combinations: replaced by the estimate for the nearest age group or by the equivalent value for England: no model convergence or events for patients in a particular age-cancer-sex combination

Presentation of results: national strategy
Joint smoothing and mapping technique



Presentation of results: local surveillance
Funnel plots for cancer survival



Conclusion and further steps

- Feasible to use flexible parametric regression models within a well defined modelling strategy to provide reliable estimates of net survival at small-area level
- Presentation of small-area survival outcomes (smoothed maps and funnel plots) provide a simple, but important quality control tool that might be helpful for both local and national health-policy makers
- Extend excess hazard models to account for the spatial (and temporal) dependency structure of the data (joint modelling)

All-Party Parliamentary Group on Cancer

All-cancers survival index by Clinical Commissioning Group

Cancer Research UK Cancer Survival Group

Manuela Quaresma
Bernard Rachet
Michel Coleman

Improving health worldwide
www.ishtm.ac.uk



All-cancers Survival Index by Clinical Commissioning Group

A tool for health policy-makers

- National – broad surveillance and health strategy
- Local – effectiveness of cancer services in each CCG

What is the All-cancers Survival Index?

- A measure of survival from **all cancers combined**
- Combines the separate survival estimates for each cancer
- Accounts for change over time in proportions by age, sex and cancer
- Allows direct interpretation of differences in survival between CCGs

Typical presentation of cancer survival statistics

One year after diagnosis

Region	Year of diagnosis															
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009		
A	67	67	68	68	69	69	70	70	71	71	71	72	73	73		
B	69	68	69	70	69	70	71	71	71	72	71	72	72	72		
C	64	65	66	67	68	69	70	71	71	72	73	73	74	75		
D	66	66	67	67	68	68	69	69	70	71	70	71	72	72		
E	67	67	68	68	68	69	69	70	70	70	71	71	72	72		
F	66	67	67	68	69	69	69	70	71	71	71	72	72	73		
G	67	67	68	69	69	70	71	71	72	73	73	73	74	74		
H	65	67	68	69	70	71	71	72	73	74	74	74	74	75		

→ 211 regions and 15 years of diagnosis
Over 3,000 estimates of survival in one table!

Cancer Survival Index by CCG

National surveillance
Geographic patterns and
time trends

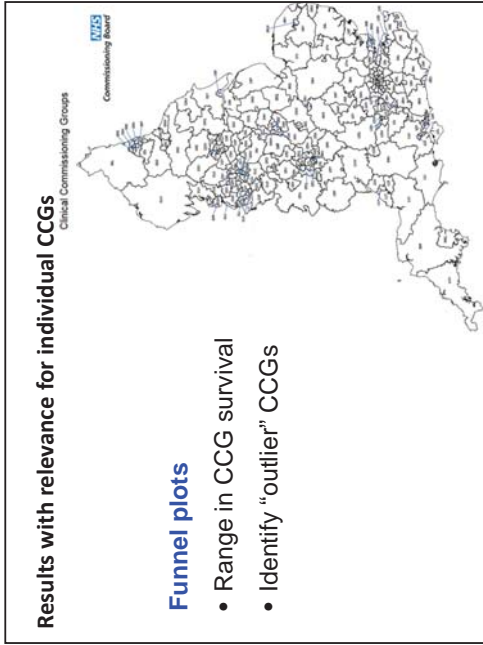
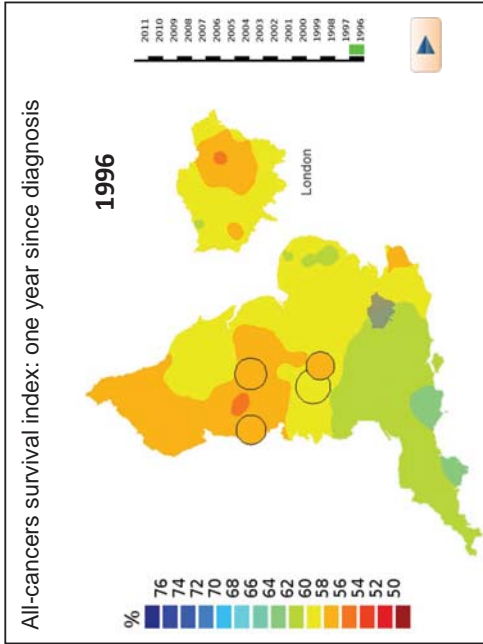
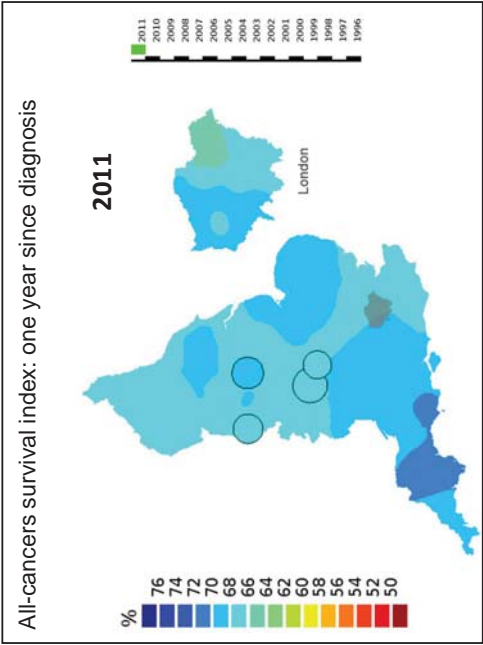
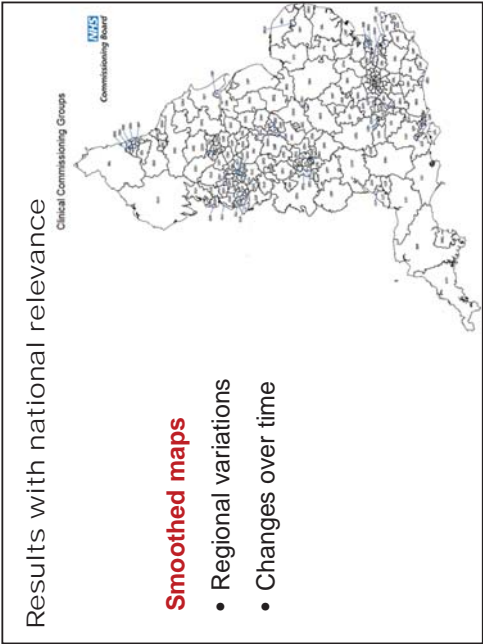


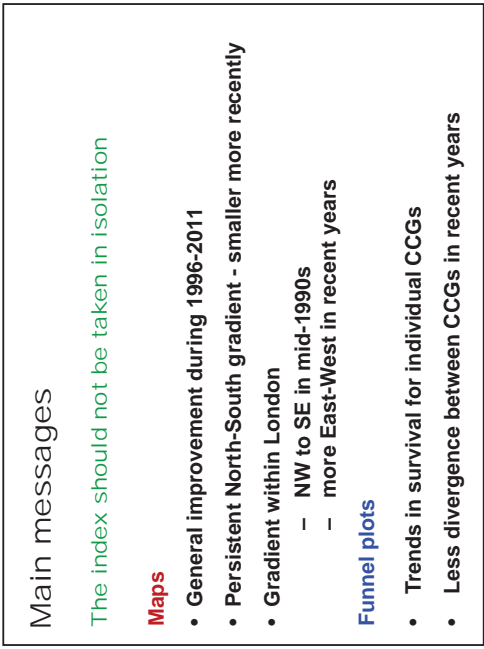
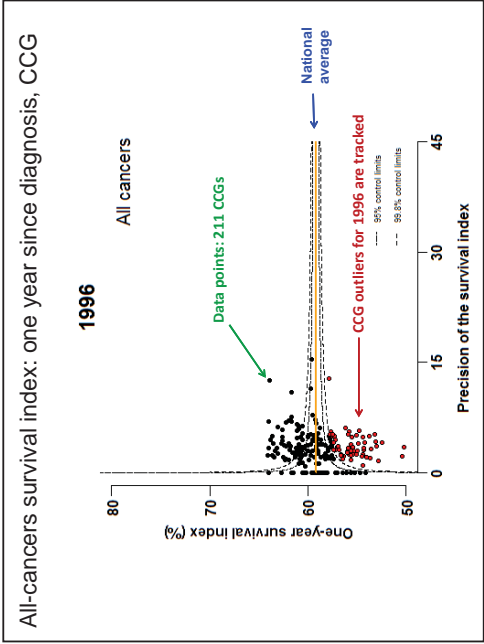
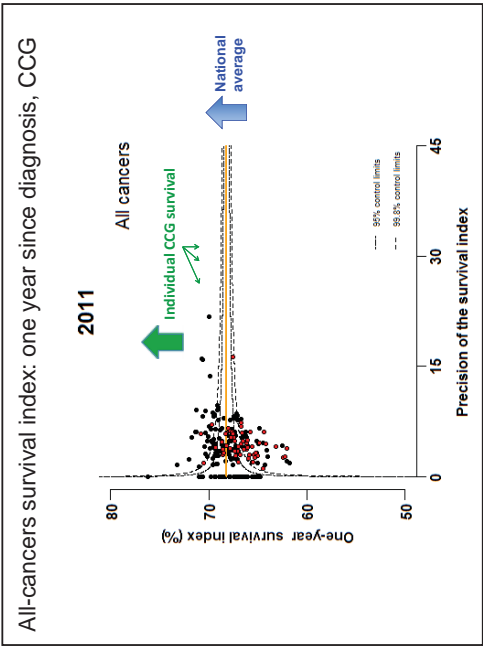
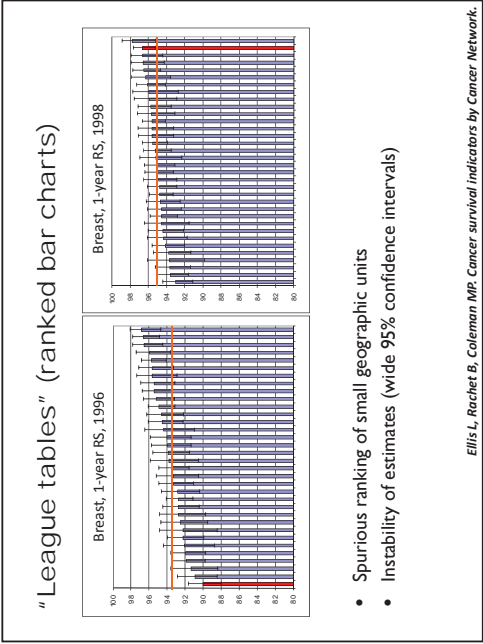
Smoothed maps

Local services
CCGs with outlying
performance



Funnel plots







Office for
National Statistics

Statistical Bulletin



Index of cancer survival for Clinical
Commissioning Groups in England;
Patients diagnosed 1996-2011 and
followed up to 2012

Coverage: England

Date: 10 December 2013

Geographical Area: Other area classification

Theme: Health and Social Care

Key findings


- The one-year net cancer survival index for England increased from 59.2% for patients diagnosed in 1996 to 68.2% in 2011
- Geographic inequalities in the one-year survival index are persistent: a clear North-South gradient existed in 1996, although this was less marked in 2011

Communication of survival patterns

- Define the users – patients, service providers, policy-makers ...
- Define the purpose – information, influence policy, monitor change
- Define the principle – statistically robust

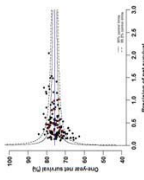
Maps

- Regional variation and changes over time
- National tool for surveillance and strategy
- Visually pleasing and accessible – to influence policy



Funnel plots

- National spread - avoids spurious ranking
- Identify "outliers" from the (moving) national average
- Local measure of effectiveness of health services






Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

Manuela Quaresma
Manuela.Quaresma@shim.ac.uk

Prof. Bernard Rachet and Prof. James Carpenter

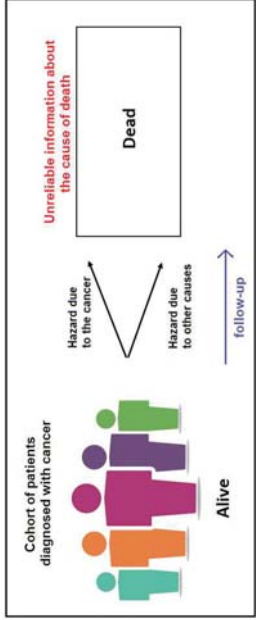
Cancer Survival Group
London School of Hygiene and Tropical Medicine
Webpage: <http://csg.shim.ac.uk> and Twitter: @csg_shim

Royal Statistical Society 2018 International Conference

Cardiff, Wales, 4 September 2018

Population-based cancer survival research

- Aim: Investigate inequalities in cancer survival
 - Estimate hazard and survival due to cancer
 - Association between prognostic factors and cancer-specific hazard



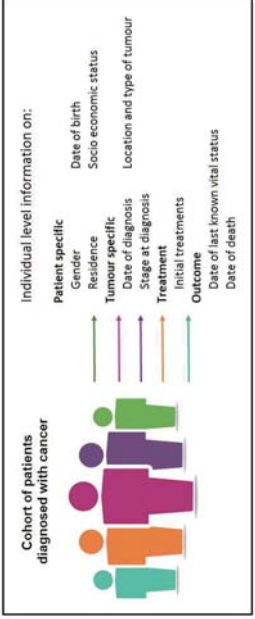
Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

4th Sep 2018

3/20

Population-based cancer survival research

- Aim: Investigate inequalities in cancer survival
 - Estimate hazard and survival due to cancer
 - Association between prognostic factors and cancer-specific hazard



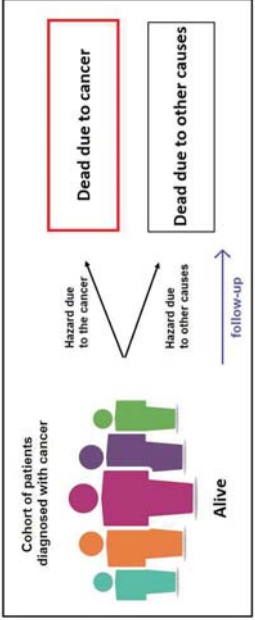
Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

4th Sep 2018

2/20

Population-based cancer survival research

- Aim: Investigate inequalities in cancer survival
 - Estimate hazard and survival due to cancer
 - Association between prognostic factors and cancer-specific hazard



Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

4th Sep 2018

4/20

Relative survival setting

- Developed for population-based cancer survival
 - Absence of recorded cause of death for each patient

Observed hazard is decomposed as

$$h(t; \mathbf{x}) = h_E(t; \mathbf{x}) + h_P(A + t; \mathbf{z})$$

- $t = (t_1, \dots, t_n)$ set of event times, and A =age at diagnosis
- $\mathbf{x} = (x_1, \dots, x_p)$ set of covariates: age, gender, deprivation, comorbidities, tumour stage, ..., and $\mathbf{z} \subset \mathbf{x}$
- $h_E(t; \mathbf{x})$ - hazard due to cancer: **excess hazard**
- $h_P(A + t; \mathbf{z})$ - hazard due to all other causes of death: **expected hazard** in the general population or **background mortality**

Main measures: excess hazard and net survival

Excess hazard

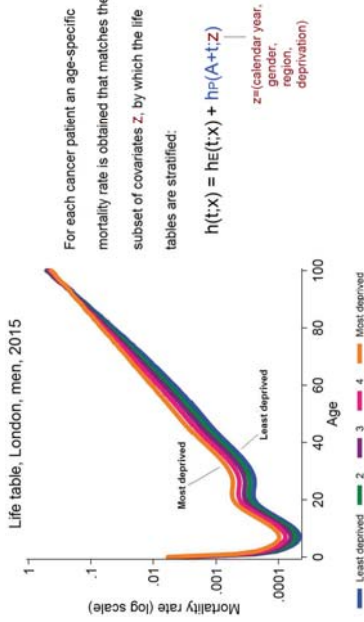
$$h(t; \mathbf{x}) = h_E(t; \mathbf{x}) + h_P(A + t; \mathbf{z})$$

Net survival

$$S_E(t; \mathbf{x}) = \exp \left\{ - \int_0^t h_E(u; \mathbf{x}) du \right\}$$

- Net survival quantifies the survival only due to the cancer, after accounting for all other causes of death in the general population
- Key health policy measure for monitoring effectiveness of the health care system in providing cancer care
- Comparing cancer management performance between countries/periods without being affected by differences in population hazard

Background mortality obtained from population life tables



Main estimators: non-parametric and parametric

- Non-parametric estimator for net survival: 'gold standard' (Perme et al., 2012)
- Parametric and semi-parametric estimators (frequentist inference):
 - Regression models on the excess hazard scale (Estève et al., 1990; Remontet et al., 2007; Charvat et al., 2016)
 - Regression models on the cumulative excess hazard scale (Lambert et al., 2009)
 - GLM formulation modelling the number of deaths (Dickman et al., 2003)

Excess hazard regression models in use

Log-likelihood

$$\sum_i \left(\delta_i \cdot \log(h_E(t_i; \mathbf{x}) + h_P(A + t_i; \mathbf{z})) - \int_0^{t_i} (h_E(u; \mathbf{x}) + h_P(A + u; \mathbf{z})) du \right)$$

$$h_E(t; \mathbf{x}) = h_{E_0}(t) \cdot \exp\left(\sum_{j \in J} \beta_j \cdot x_j + \sum_{k \in K} f_k(x_k) + \sum_{l \in L} g(l) \cdot x_l\right)$$

- $h_{E_0}(t)$ - logarithm of the baseline excess hazard commonly modelled using flexible functions (restricted cubic splines* or B-splines*)
- Variables in **set J** modelled with a linear effect
- Non-linear effects in **set K** modelled using flexible functions*
- Non-proportional effects in **set L** modelled including interaction between covariates and time

Added complexity: likelihood formulation with no closed-form expression

Common solution: use numerical integration rules

Choice of splines: Low-Rank Thin Plate splines

- Likelihood function remains tractable, avoiding numerical integration
- First-order polynomials - penalised splines
- Simple yet enough flexibility to capture the shapes of excess hazards

Piecewise linear log-baseline excess hazard model:

Given a partition of the follow-up time range as $0 = \tilde{t}_0 < \tilde{t}_1 < \dots < \tilde{t}_K = \infty$

$$\log(h_{E_0}(t; \alpha^*)) = \alpha_0^* + \alpha_1^* t + \sum_{k=2}^K \alpha_k^* (|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|)$$

Implementation involves a series of transformations to the spline parameters α^* , as well as constructing a time design matrix and a penalty transformation matrix (Crainiceanu et al. 2005 and Murray et al., 2016)

Proposed excess hazard model

Aims

- Model that could accommodate hierarchical data structures
 - ▶ cancer patients clustered within health geographies (small areas)
 - ▶ cancer patients clustered within hospital of treatment
- Log-baseline excess hazard modelled using flexible function
- Include non-linear and non-proportional effects
- Inference within Bayesian framework suitable for these model specifications
- **First step:** introduce flexible Bayesian model for the excess hazard (without random effects)

Illustration: using population-based cancer data

- **Data:** All adult men diagnosed with colon cancer during 2009 in London, and followed-up until the 31st December 2015.
- **Variables available for analysis:**
 - ▶ full dates of diagnosis, last follow-up and death
 - ▶ vital status indicator (dead or alive at the end of follow-up)
 - ▶ age at diagnosis (15-99 years)
 - ▶ deprivation quintile (Index of Multiple Deprivation - income domain)
 - ▶ region of residence (32 regions), hospital of treatment (38 hospitals)
- **Background mortality:** Life tables for England stratified by calendar year (2009-2015), age, gender, deprivation quintile and region of residence

Illustration: Model set-up

Age at diagnosis (A) and deprivation quintile (dep) as main effects:

$$\begin{aligned} \log(h_E(t|\alpha; \beta, \gamma)) = & (\alpha_{0,0} + \alpha_{1,0}A) + (\alpha_{0,1} + \alpha_{1,1}A)t \\ & + \sum_{k=2}^K (\alpha_{0,k} + \alpha_{1,k}A)(|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \quad \text{[part 1]} \\ & + \beta_1^*(A - \bar{A}) + \sum_{j=2}^J \beta_j^* (|A - \tilde{A}_{j-1}|^9 - |\bar{A} - \tilde{A}_{j-1}|^9) \quad \text{[part 2]} \\ & + \gamma * dep \quad \text{[part 3]} \end{aligned}$$

- **part 1** spline modelling the baseline log-excess hazard using 4 partitions of the observed follow-up time, and incorporating the time-dependent effect of age at diagnosis.
- **part 2** spline modelling the non-linear effect of age at diagnosis using 3 partitions (J=3) of the observed age range.
- **part 3** formulates the linear and proportional effect of deprivation.

Illustration: Posterior Excess Hazard Ratios (mean)

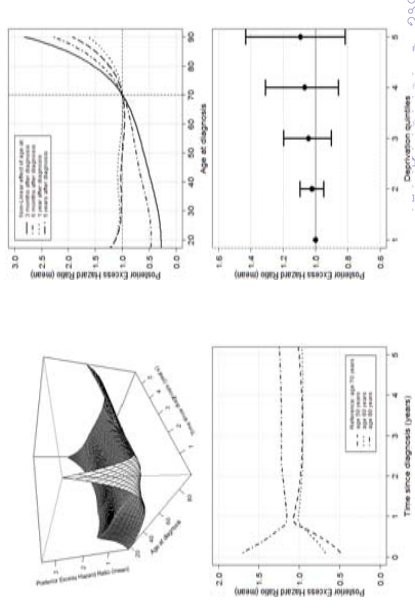


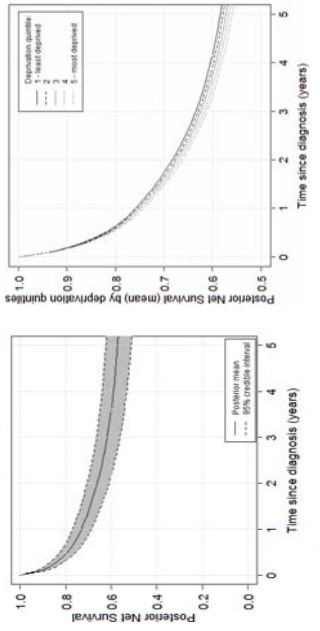
Illustration: Bayesian estimation

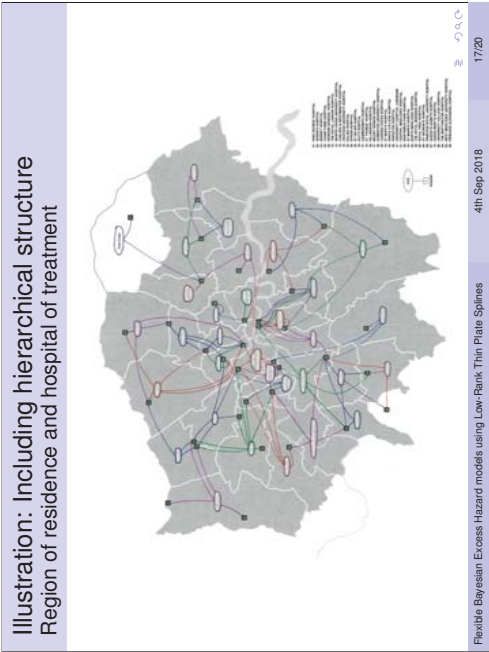
- Vague prior distributions used for the parameters:
 - for the baseline log-excess hazard:

$$\alpha_0 \sim N(0, 10^4), \alpha_1 \sim N(0, 10^4)$$

$$\alpha_k | \sigma_{\alpha_k} \stackrel{iid}{\sim} N(0, \sigma_{\alpha_k}^2), \text{ for } k=2, \dots, K \text{ and } \sigma_{\alpha_k} \sim U(0.01, 100)$$
- Model fitted in JAGS accessed via R2JAGS
- 30,000 MCMC samples from each posterior distribution
- Examination of trace and density plots did not indicate any convergence issues
- MCMC samples were saved and posterior distributions for excess hazard and net survival were derived in a post-estimation procedure

Illustration: Posterior Net survival (mean)





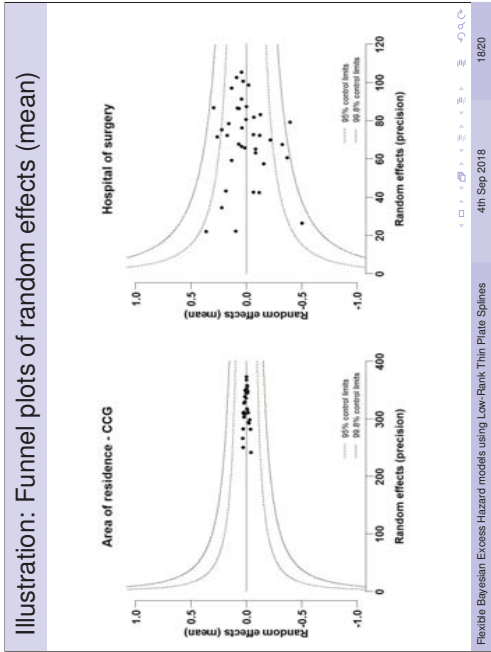
References

- J. Esteve, E. Benhamou, M. Croasdale, and L. Raymond. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 1990.
- L. Remontet, N. Bossard, A. Belot, J. Esteve, and FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, 2007.
- H. Charvat, L. Remontet, N. Bossard, L. Roche, O. Dejaridin, B. Rachet, G. Launoy, A. Belot. CENSUR Working Survival Group. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 2016.
- P. W. Dickman, A. Sloggett, M. Hills, and T. Hakulinen. Regression models for relative survival. *Statistics in Medicine*, 2004.
- P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *Stata Journal*, 2009.
- Maja Pohar Perme, Janez Stare, and Jacques Esteve. On estimation in relative survival. *Biometrics*, 2012.
- C. M. Crainiceanu, D. Ruppert, and M. P. Wand. Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software*, 2005.
- T.A. Murray, B.P. Hobbs, D.J. Sargent, and B.P. Carlin. Flexible bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian Analysis*, 2016.

Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

4th Sep 2018

19/20



Thank you for your attention

Manuela.Quaresma@lsthm.ac.uk

Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate Splines

4th Sep 2018

20/20

Appendix C

Ethical approvals

Ethical approvals were obtained from the Ethics Committee of the London School of Hygiene & Tropical Medicine (LSHTM; number 5192) and the NHS South East Research Ethics Committee (07/MRE01/52). Ethical approval to analyse the data in this PhD was obtained from the ONS Medical Research Service (MR1101), from the statutory Patient Information Advisory Group (PIAG; now the Ethics and Confidentiality Committee of the National Information Governance Board) under Section 61 of the Health and Social Care Act 2001 (PIAG 1-05(c)/2007) and from the Security and Confidentiality Advisory Group (SCAG (HES) AG/65/5/b).

Bibliography

- [1] Quaresma M, Coleman MP, and Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *The Lancet*, 385:1206–1218, 2015.
- [2] Quaresma M, Whitehead S, Bannister N, Coleman MP, and Rachet B. Index of cancer survival for Clinical Commissioning Groups in England; Patients diagnosed 1996-2011 and followed up to 2012. Technical report, Office for National Statistics, Newport, UK, 2013.
- [3] Quaresma M, Drummond R, Rowlands S, Brown P, Bannister N, Coleman MP, and Rachet B. Index of cancer survival for Clinical Commissioning Groups in England: Adults diagnosed 1997-2012 and followed up to 2013. Technical report, Office for National Statistics, Newport, UK, 2014.
- [4] Quaresma M, Nash E., Bannister N., Coleman MP., and Rachet B. Index of cancer survival for Clinical Commissioning Groups in England: Adults diagnosed 1998-2013 and followed up to 2014. Technical report, Office for National Statistics, Newport, UK, 2016.
- [5] Quaresma M, Jenkins J, Bannister N, Murphy R, Kaur J, Peet M, Magadi W, Coleman MP, and Rachet B. Index of cancer survival for Clinical Commissioning Groups in England: adults diagnosed 1999-2014 and followed up to 2015. Technical report, Office for National Statistics, Newport, UK, 2016.

- [6] Quaresma M, Coleman MP, and Rachet. Funnel plots for population-based cancer survival: principles, methods and applications. *Statistics in Medicine*, 2013. doi: 10.1002/sim.5953.
- [7] International Agency for Research on Cancer. Cancer today (GLOBOCAN), 2018. URL <http://gco.iarc.fr/today/home>.
- [8] Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, Bonaventure A, Valkov M, Johnson CJ, Estève J, Ogunbiyi OJ, Azevedo e Silva G, Chen WQ, Eser S, Engholm G, Siller CA, Monnereau A, Woods RR, Visser O, Lim GH, Aitken J, Weir HK, Coleman MP, and the Concord Working Group. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37513025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, 2018.
- [9] Office for National Statistics. Cancer registration statistics, England: first release, 2016.
- [10] Bray F. The burden of cancer in Europe. In Coleman MP, Alexe DM, Albrecht T, and McKee CM, editors, *Responding to the challenge of cancer in Europe*, book chapter 2, pages 7–40. Institute of Public Health of the Republic of Slovenia, 2008.
- [11] dos Santos Silva I. *Cancer epidemiology: principles and methods*. International Agency for Research on Cancer, Lyon, 1999.
- [12] Curado MP, Edwards BK, Shin HR, Storm HH, Ferlay J, Heanue M, and Boyle P. Cancer incidence in five continents. vol. IX. IARC scientific publications no. 160. Serial (Book, Monograph), 2007.
- [13] Coleman MP and Estève J. Trends in cancer incidence, survival and mortality. In Husband JE and Reznik RH, editors, *Imaging in Oncology*, volume 1, book chapter 2, pages 11–22. Medical Media, 1998.
- [14] Ellis L, Woods LM, Estève J, Eloranta S, Coleman MP, and Rachet B. Cancer incidence, survival and mortality: explaining the concepts. *International Journal of Cancer*, 135:1774–1782, 2014.

- [15] Niksic M, Rachet B, Duffy SW, Quaresma M, Møller H, and Forbes LJ. Is cancer survival associated with cancer symptom awareness and barriers to seeking medical help in England? An ecological study. *British Journal of Cancer*, 115(7):876–886, 2016.
- [16] Institute of Medicine of the National Academies. *The unequal burden of cancer*. National Academy Press, Washington, D.C., 1999.
- [17] Richards MA. The national awareness and early diagnosis initiative in England: assembling the evidence. *British Journal of Cancer*, 101 (Suppl. 2):1–4, 2009.
- [18] Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, Brenner H, and Estève J. Trends and socio-economic inequalities in cancer survival in England and Wales up to 2001. *British Journal of Cancer*, 90:1367–1373, 2004.
- [19] Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, Steward JA, Brenner H, Estève J, Sullivan R, and Coleman MP. Cancer survival in England and Wales at the end of the 20th century. *British Journal of Cancer*, 99 (Suppl. 1):2–10, 2008.
- [20] World Health Organisation. *National Cancer Control Programmes*. Geneva, 2002.
- [21] Union for International Cancer Control. World cancer declaration, 2013. URL <http://www.uicc.org/world-cancer-declaration>.
- [22] World Health Organisation. The world health report. Health Systems: Improving performance. Technical report, 2000.
- [23] Mooney G. The demand for effectiveness, efficiency and equity of health care. *Theoretical Medicine*, 10:195–205, 1989.
- [24] Mooney G. Equity in health care: Confronting the confusion. *Effective Health Care*, 1:179–185, 1983.
- [25] Mooney G. What does equity in health mean? *World Health Statistics*, 40:296–303, 1987.
- [26] Sassi F, Le Grand J, and Archard L. Equity versus efficiency: a dilemma for the NHS. *British Medical Journal*, 323:762–3, 2001.

- [27] National Health Service (NHS). Principles and values that guide the NHS, 2018. URL <https://www.nhs.uk/NHSEngland/thenhs/about/Pages/nhscoreprinciples.aspx>.
- [28] Working Group on Inequalities in Health. Inequalities in health: report of a research working group [the black report]. Report, 1980.
- [29] Coleman MP. Cancer survival: global surveillance will stimulate health policy and improve equity. *The Lancet*, 383(9916):564–573, 2014. doi: [https://doi.org/10.1016/S0140-6736\(13\)62225-4](https://doi.org/10.1016/S0140-6736(13)62225-4).
- [30] Department of Health. Health and social care bill, 2011. URL <http://www.dh.gov.uk/en/Publicationsandstatistics/Legislation/Actsandbills/HealthandSocialCareBill2011/index.htm>.
- [31] Department of Health and Parliament Social Care. Health and social care act, 2012. URL <http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted>.
- [32] Walshe K, Smith J, Dixon J, Edwards N, Hunter DJ, Mays N, Normand C, and Robinson R. Primary Care Trusts. Premature reorganisation, with mergers, may be harmful. *British Medical Journal*, 329(7471):871–872, 2004.
- [33] Office for National Statistics. English health geographies, 2012. URL <http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/health/english-health-geography/index.html>.
- [34] Ellis L, Rachet B, and Coleman MP. Cancer survival indicators by Cancer Network: a methodological perspective. *Health Statistics Quarterly*, Winter:36–41, 2007.
- [35] Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, and Estève J. Survival of cancer patients in Europe: the EURO CARE study (IARC Scientific Publications No. 132). 1995.
- [36] Capocaccia R, Gatta G, Chessa E, Valente F, and EURO CARE Working Group. The EURO CARE-2 study. In Berrino F, Capocaccia R, Estève J, Gatta G, Hakulinen T, Micheli A, Sant M, and Verdecchia A, editors, *Survival of cancer patients in Europe: the EURO CARE-2 study*. (IARC Scientific Publications No. 151). 1999.

- [37] Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R, and EURO-CARE Working Group. EURO-CARE-4. Survival of cancer patients diagnosed in 1995-1999: results and commentary. *European Journal of Cancer*, 45:931–991, 2009.
- [38] De Angelis R, Sant M, Coleman MP, Francisci S, Baili P, Pierannunzio D, Trama A, Visser O, Brenner H, Ardanaz E, Bielska-Lasota M, Engholm G, Nennecke A, Siesling S, Berrino F, Capocaccia R, and the EURO-CARE-5 Working Group. Cancer survival in Europe 1999-2007 by country and age: results of EURO-CARE-5 - a population-based study. *Lancet Oncology*, 15:23–34, 2014.
- [39] Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, Marrett L, Gjerstorff ML, Johannesen TB, Adolfsson J, Lambe M, Lawrence G, Meechan D, Morris EJ, Middleton R, Steward J, Richards MA, and ICBP Module 1 Working Group. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *The Lancet*, 377:127–138, 2011.
- [40] Farrow DC, Samet JM, and Hunt WC. Regional variation in survival following the diagnosis of cancer. *Journal of Clinical Epidemiology*, 49(8):843–847, 1996.
- [41] Rachet B, Riga M, Mitry E, Romanengo M, Quinn MJ, Cooper N, and Coleman MP. Geographical comparisons of cancer survival indicators. *Health Statistics Quarterly*, 22:5–13, 2004.
- [42] Coleman MP, Babb P, Damiecki P, Grosclaude PC, Honjo S, Jones J, Knerer G, Pitard A, Quinn MJ, Sloggett A, and De Stavola BL. *Cancer survival trends in England and Wales 1971-1995: deprivation and NHS Region. (Studies on Medical and Population Subjects No. 61)*. The Stationery Office, London, 1999.
- [43] Coleman MP, Babb P, Quinn MJ, Sloggett A, and De Stavola BL. Socio-economic inequalities in cancer survival in England and Wales. *Cancer*, 91:208–216, 2001.

- [44] Mullee MA, BL De Stavola, Romanengo M, and Coleman MP. Geographical variation in breast cancer survival rates for women diagnosed in England between 1992 and 1994. *British Journal of Cancer*, 90(11):2153–2156, 2004.
- [45] Yuen J, Haybittle J, and Machin D. Geographical variation in the standardised years of potential life lost ration in women dying from malignancies of the breast in England and Wales. *British Journal of Cancer*, 75:1069–1074, 1997.
- [46] Rachet B, Ellis L, Maringe C, Nur U, Chu T, Quaresma M, Shah A, Walters S, Woods LM, Forman D, and Coleman MP. Socioeconomic inequalities in cancer survival in England after the NHS Cancer Plan. *British Journal of Cancer*, 103:446–453, 2010.
- [47] Walters S, Quaresma M, Coleman MP, Gordon E, Forman D, and Rachet B. Geographical variation in cancer survival in England, 1991-2006: an analysis by Cancer Network. *Journal of Epidemiology and Community Health*, 2011.
- [48] Nur U, Coleman MP, Gordon E, Jakomis N, Carrigan C, and Rachet B. Cancer survival by Cancer Network in England - patients diagnosed 1996-2009 and followed up to 2010. Technical report, Office for National Statistics, 2011.
- [49] Ellis L, Coleman MP, and Rachet B. How many deaths would be avoidable if socioeconomic inequalities in cancer survival in England were eliminated? a national population-based study, 1996-2006. *European Journal of Cancer*, 48:270–278, 2012.
- [50] Lyratzopoulos G, Barbiere JM, Rachet B, Baum M, Thompson MR, and Coleman MP. Changes over time in socioeconomic inequalities in breast and rectal cancer survival in England and Wales over a 32-period (1973-2004): the potential role of health care. *Annals of Oncology*, 22:1661–1666, 2011.
- [51] Woods LM, Sasieni P, and Rachet B. Screening mammography and socioeconomic inequalities in breast cancer survival. *Annals of Oncology*, 23:285–286, 2012.
- [52] Shack LG, Rachet B, Brewster DH, and Coleman MP. Socioeconomic inequalities in cancer survival in Scotland 1986-2000. *British Journal of Cancer*, 97:999–1004, 2007.

- [53] Exarchakou A, Rachet B, Belot A, Maringe C, and Coleman MP. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *British Medical Journal*, 2018.
- [54] Nur U, Lyratzopoulos G, Rachet B, and Coleman MP. The impact of age at diagnosis on socioeconomic inequalities in adult cancer survival in England. *Cancer Epidemiology*, 39:641–649, 2015.
- [55] Department of Health. Baroness Jay says breast cancer services are major priority. Report, 1997.
- [56] Baron J. Cancer services. *Commons Hansard [Parliamentary Record]*, 1 March: c320–c331, 2006.
- [57] All Party Parliamentary Group on Cancer. All Party Parliamentary Group on Cancer's inquiry into inequalities of cancer. Report, 2009.
- [58] A Report by the Expert Advisory Group on Cancer to the Chief Medical Officers of England and Wales. A policy framework for commissioning cancer services (the Calman-Hine report). London: Department of Health, 1995.
- [59] House of Commons Committee of Public Accounts. Tackling cancer in England: saving more lives. Report, 2005.
- [60] Department of Health. Cancer reform strategy. Report, 2007.
- [61] House of Commons Committee of Public Accounts. Delivering the cancer reform strategy. hc667, session 2010-11. Report, 2011.
- [62] Department of Health. The NHS Cancer Plan. Report, 2000.
- [63] House of Commons. House of Commons Minutes of Evidence "Tackling cancer in England: saving more lives". Oral evidence given by Sir Nigel Crisp KCB and Professor Mike Richards CBE. Generic, 2004.
- [64] Department of Health. Cancer Reform Strategy: Equality Impact Assessment, 2007.
- [65] Department of Health. Cancer Reform Strategy: achieving local implementation - second annual report, 2009.

- [66] National Audit Office. Delivering the Cancer Reform Strategy. HC 568, Session 2010-2011, 2010.
- [67] Department of Health. NHS Long Term Plan for Cancer. Report, January 2019.
- [68] Independent Cancer Taskforce. Achieving world-class cancer outcomes. A strategy for England 2015-2020. Report, 2015.
- [69] All-Party Parliamentary Group on Cancer. All Party Parliamentary Group on Cancer Inquiry: Progress of the England Cancer Strategy: Delivering outcomes by 2020? Report, 2017. URL https://www.macmillan.org.uk/_images/progress-of-the-england-cancer-strategy-delivering-outcomes-by-2020_tcm9-321006.pdf.
- [70] Estève J, Benhamou E, and Raymond L. *Statistical methods in cancer research, volume IV. Descriptive epidemiology. (IARC Scientific Publications No. 128)*. International Agency for Research on Cancer, Lyon, 1994.
- [71] Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics, 2002.
- [72] Collett D. *Modelling Survival Data in Medical Research*. Chapman and Hall, second edition, 2003.
- [73] Schaffar R, Rachet B, Belot A, and Woods LM. Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data. *European Journal of Cancer*, 72:78–83, 2017.
- [74] Percy CL, Stanek E, and Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*, 71: 242–250, 1981.
- [75] Ashworth TG. Inadequacy of death certification: proposal for change. *Journal of Clinical Pathology*, 44:265–268, 1991.
- [76] Ranganathan P and Pramesh CS. Censoring in survival analysis: Potential for bias. *Perspectives in Clinical Research*, 3(1):40, 2012.

- [77] Pohar-Perme M, Stare J, and Estève J. On estimation in relative survival. *Biometrics*, 68(1):113–120, 2012.
- [78] Rebolj KA and Pohar-Perme M. Informative censoring in relative survival. *Statistics*, 32(27):4791–802, 2013.
- [79] Berkson J and Gage RP. Calculation of survival rates for cancer. *Proceedings Staff Meetings Mayo Clinic*, 25:270–286, 1950.
- [80] Cutler SJ and Ederer F. Maximum utilisation of the life table method in analyzing survival. *Journal of Chronic Diseases*, 8:699–712, 1958.
- [81] Estève J, Benhamou E, Croasdale M, and Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 9: 529–538, 1990.
- [82] Ederer F and Heise H. Instructions to IBM 650 programmers in processing survival computations, methodological note no. 10, end results evaluation section. Technical report, National Cancer Institute, Bethesda MD, 1959.
- [83] Ederer F, Axtell LM, and Cutler SJ. The relative survival: a statistical methodology. *National Cancer Institute Monograph series*, 6:101–121, 1961.
- [84] Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38:933–942, 1982.
- [85] Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *Journal American Statistical Association*, 53:457–481, 1958.
- [86] Gehan EA. Estimating survival functions from the life table. *Journal of Chronic Diseases*, 21(9):629–644, 1969.
- [87] Chiang CL. A stochastic study of the life table and its applications. II. Sample variance of the observed expectation of life and other biometric functions. *Human Biology*, 32:221–238, 1960.
- [88] Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, Iwaz J, Remontet L, and Bossard N. Cancer net survival on registry data: Use of the new unbiased

- Pohar-Perme estimator and magnitude of the bias with the classical methods. *International Journal of Cancer*, 132:2359–2369, 2013.
- [89] Pohar-Perme M, Estève J, and Rachet B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer*, 16(933), 2016.
- [90] Pohar Perme M and Pavlič K. Nonparametric relative survival analysis with the *r* package relsurv. *Journal of Statistical Software*, 87(8), 2018.
- [91] Clerc-Urmès I, Grzebyk M, and Hédelin G. Net survival estimation with stns. *The Stata Journal*, 14(1):87–102, 2014.
- [92] Cox DR. Regression models and life-tables. *Journal Royal Statistical Society, Series B*, 34:187–200, 1972.
- [93] Abrahamowicz M, MacKenzie T, and Esdaile JM. Time-dependent hazard ratio: modelling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, 91(436):1432–1439, 1996.
- [94] Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879, 1992.
- [95] Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87: 942–951, 1992.
- [96] Royston P and Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15): 2175–2197, 2002.
- [97] Royston P and Lambert PC. *Flexible Parametric Survival Analysis using Stata: Beyond the Cox Model*. Stata Press, 2011.
- [98] Abrahamowicz M and MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, 26(2): 392–408, 2007.

- [99] Remontet L, Bossard N, Belot A, Estève J, and FRANCIM. An overall strategy based on regression models to estimate relative survival and models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, 26:2214–2228, 2007.
- [100] Lambert PC and Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal*, 9:265–290, 2009.
- [101] Dickman PW, Sloggett A, Hills M, and Hakulinen T. Regression models for relative survival. *Statistics in Medicine*, 23:51–64, 2004.
- [102] Lambert PC, Smith LK, Jones DR, and Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, 24(24):3871–3885, 2005.
- [103] Stata statistical software: Release 15. *StataCorp LLC*, 2017. URL <https://www.stata.com/>.
- [104] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2019. URL <http://www.R-project.org/>.
- [105] Rachet B, Maringe C, Woods LM, Ellis L, Spika D, and Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health*, 15:1240, 2015.
- [106] Bolard P, Quantin C, Abrahamowicz M, Estève J, Giorgi R, Chadha-Boreham H, Binquet C, and Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention*, 7(3):113–122, 2002.
- [107] Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, Launoy G, Belot A, and CENSUR Working Survival Group. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 35(18):3066–84, 2016.
- [108] Remontet L, Uhry Z, Bossard N, Iwaz J, Belot A, Danieli C, Charvat H, Roche L, and the CENSUR Working Survival Group. Flexible and structured survival model

- for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis. *Statistical Methods Medical Research*, 28(8):2368–2384, 2019.
- [109] Fauvernier M, Roche L, Uhry Z, Tron L, Bossard N, and Remontet L. Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society, series C*, 68(5):1233–1257, 2019.
- [110] Cancer Research UK Cancer Survival Group. strel computer program version 5.8 for cancer survival analysis. Electronic Citation, 2010.
- [111] Bower H, Crowther MJ, and Lambert PC. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. *The Stata Journal*, 16(4):989–1012, 2016.
- [112] Charvat H and Belot A. Mixed Effect Excess Hazard Models (mexhaz). R package, 2019.
- [113] Durrleman S and Simon R. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.
- [114] Nelson CP, Lambert PC, Squire IB, and Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26:5486–5498, 2007.
- [115] Breslow NE and Day NE. *Statistical methods in cancer research, volume II. The design and analysis of cohort studies. (IARC Scientific Publications No. 82)*. International Agency for Research on Cancer, Lyon, 1987.
- [116] Dickman PW and Coviello E. Estimating and modeling relative survival. *The Stata Journal*, 15(1):186–215, 2015.
- [117] Danieli C, Remontet L, Bossard N, Roche L, and Belot A. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*, 31(8):775–86, 2012.

- [118] Jensen OM, Parkin DM, MacLennan R, Muir CS, and Skeet RG. Cancer registration: principles and methods. (IARC scientific publication no. 95). Serial (Book, Monograph), 1991.
- [119] Tyczynski JE, Démaret E, and Parkin DM. Standards and guidelines for cancer registration in Europe. the ENCR recommendations. IARC technical publication no. 40. Report, 2003.
- [120] Bray F, Znaor A, Cueva P, Korir A, Swaminathan R, Ullrich A, Wang SA, and Parkin DM. Planning and developing population-based cancer registration in low- and middle-income settings. IARC technical publication no. 43, International Agency for Research on Cancer, Lyon, France, 2014.
- [121] Muir CS. The cancer registry in cancer control: an overview. *Archives of Geshwulstforsch*, 54:491–497, 1984.
- [122] Li R, Abela L, Moore J, Woods LM, Nur U, Rachet B, Allemani C, and Coleman MP. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiology*, 38:314–320, 2014.
- [123] World Health Organisation. *International statistical classification of diseases and related health problems. Tenth revision*. WHO, Geneva, 1994.
- [124] World Health Organisation. *International Classification of Diseases for Oncology (ICD-O)*. World Health Organisation, Geneva, 1976.
- [125] V. Carstairs. Deprivation indices: their interpretation and use in relation to health. *J. Epidemiol. Comm. Hlth.*, 49 (Suppl 2):3–8, 1995.
- [126] R. Morris and V. Carstairs. Which deprivation? a comparison of selected deprivation indexes. *J. Publ. Hlth. Med.*, 13:318–326, 1991.
- [127] Department of the Environment Transport and the Regions. Measuring multiple deprivation at the small area level: the indices of deprivation 2000. Report, 2000.
- [128] Neighbourhood Renewal Unit. The english indices of deprivation 2004 (revised). Report, 2004.

- [129] Ministry of Housing, Communities and Local Government. English indices of deprivation, 2015. URL <https://www.gov.uk/government/collections/english-indices-of-deprivation>.
- [130] L. M. Woods, B. Rachet, and M. P. Coleman. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. *Br. J. Cancer*, 92:1279–1282, 2005.
- [131] Diez Roux AV. Investigating neighborhood and area-effects on health. *American Journal of Public Health*, 91(11):1783–1789, 2001.
- [132] ONS Geography. A guide to ONS Geography Postcode Products, 2016. URL <https://geoportal.statistics.gov.uk/datasets/a-guide-to-ons-geography-postcode-products>.
- [133] NHS Digital. Hospital Episode Statistics (HES), 2018. URL <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics>.
- [134] Clinical Audit and Registries Management Service. National Bowel Cancer Audit (NBOCA), 2018. URL <https://www.nboca.org.uk/>.
- [135] Cornish JA, Tekkis PP, Tan E, Tilney HS, Thompson MR, and Smith JJ. The national bowel cancer audit project: The impact of organisational structure on outcome in operative bowel cancer within the united kingdom. *Surgical Oncology*, 20(2):e72–e77, 2011.
- [136] NHS Digital. National Bowel Cancer Audit, 2018. URL <https://digital.nhs.uk/data-and-information/clinical-audits-and-registries/national-bowel-cancer-audit>.
- [137] Cancer Survival Group. Life tables for England and Wales by sex, calendar period, region and deprivation, 2019. URL <https://csg.lshtm.ac.uk/life-tables/>.
- [138] Ferlay J, Parkin DM, Curado MP, Bray F, Edwards BK, Shin HR, and Forman D. Cancer incidence in five continents, volumes I to IX: IARC CancerBase No. 9, 2010. URL <http://ci5.iarc.fr>.

- [139] Office for National Statistics. Cancer survival statistical bulletins - quality and methodology information. Technical report, Office for National Statistics, 2017.
- [140] Magadi W, Rachet B, Quaresma M, Exarchakou A, Belot A, Nash E, Bannister N, Rowlands S, and Coleman MP. Childhood cancer survival in England, children diagnosed from 1990 to 2009 and followed up to 2014. Technical report, Office for National Statistics, 2016.
- [141] Magadi W, Exarchakou A, Rachet B, Coleman MP, Jenkins J, Bannister N, Murphy R, and Rowlands S. Cancer survival in England: patients diagnosed between 2010 and 2014 and followed up to 2015. Technical report, Office for National Statistics, 2016.
- [142] Belot A, Coleman MP, Magadi W, Kaur J, Peet M, Rowlands S, Bannister N, and Rachet B. Geographic patterns of cancer survival in England for adults diagnosed 2003 to 2010 and followed up to 2015. Technical report, Office for National Statistics, 2017.
- [143] Cancer Survival Group, London School of Hygiene and Tropical Medicine, 2019. URL <https://csg.lshtm.ac.uk/>.
- [144] Rachet B, Maringe C, Nur U, Quaresma M, Shah A, Woods LM, Ellis L, Walters S, Forman D, Steward JA, and Coleman MP. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncology*, 10:351–369, 2009.
- [145] Corazziari I, Quinn MJ, and Capocaccia R. Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40:2307–2316, 2004.
- [146] Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, and Inoue M. Age standardization of rates: a new WHO standard. GPE Discussion Paper Series: No.31. EIP/GPE/EBD World Health Organization.
- [147] EUROSTAT. Revision of the European standard population. Technical report, 2013. URL <https://ec.europa.eu/eurostat/en>.

- [148] Black RJ and Bashir SA. World standard cancer patient populations: a resource for comparative analysis of survival data. In Sankaranarayanan R, Black RJ, and Parkin DM, editors, *Cancer survival in developing countries (IARC Scientific Publications No. 145)*, pages 9–11. International Agency for Research on Cancer, 1998.
- [149] Capocaccia R, Gatta G, Roazzi P, Carrani E, Santaquilani M, De Angelis R, Tavilla A, and EUROCARE Working Group. The EUROCARE-3 database: methodology of data collection, standardisation, quality control and statistical analysis. *Annals of Oncology*, 14 (Suppl. 5):14–27, 2003.
- [150] Cancer Registration Statistics, England: 2011. Technical report, Office for National Statistics, Newport, UK, 2011.
- [151] Harrell F. *Regression Modeling Strategies*. Springer, 2001.
- [152] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- [153] Lambert P. RCGEN: Stata module to generate restricted cubic splines and their derivatives, 2008. URL <https://pclambert.net/software/rcsge/>.
- [154] Casella G and Berger RL. *Statistical Inference*. Cengage Learning, 2nd edition edition, 2008.
- [155] Cancer Research UK. Beating Cancer sooner. Our Research strategy. Technical report, 2014. URL https://www.cancerresearchuk.org/sites/default/files/cruk_research_strategy.pdf.
- [156] Jones G. Why are cancer rates increasing?, 2015. URL <https://scienceblog.cancerresearchuk.org/2015/02/04/why-are-cancer-rates-increasing/>.
- [157] All-Party Parliamentary Group on Cancer. One Year Cancer Survival Rates: Measuring Progress, 2015. URL <https://www.macmillan.org.uk/documents/campaigns/appgc-measuringprogress.pdf>.
- [158] John Baron. Chairman of the All Party Parliamentary Group on Cancer (APPGC), 2009–2018. URL <http://www.johnbaron.co.uk/cancer.html>.

- [159] John Baron MP. NHS transparency on cancer survival rates will be transformational, 2014. URL <http://www.conservativehome.com/platform/2014/12/john-baron-mp-nhs-transparency-on-cancer-survival-rates-/will-be-transformational.html>.
- [160] Johnson CJ, Weir HK, Mariotto A, Wilson R, and Nishri D. Construction of a North American Cancer Survival Index to Measure Progress of Cancer Control Efforts. *Preventing Chronic Disease*, 14, 2017.
- [161] Bell S, Hoskins RE, Pickle LW, and Wartenberg D. Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. *International Journal of Health Geographics*, 5(49), 2006.
- [162] Tripathy JP, Bhatnagar A, Shewade HD, Kumar AMV, Zachariah R, and Harries AD. Ten tips to improve the visibility and dissemination of research for policy makers and practitioners. *Public Health Action*, 7(1):10–14, 2017.
- [163] Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T, Micheli A, Sant M, Weir HK, Elwood JM, Tsukuma H, Koifman S, Azevedo e Silva Gand Francisci S, Santaquilani M, Verdecchia A, Storm HH, Young JL, and CONCORD Working Group. Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncology*, 9: 730–756, 2008.
- [164] Berrino F, Capocaccia R, Coleman MP, Estève J, Gatta G, Hakulinen T, Micheli A, Sant M, and Verdecchia A. EUROCare-3: the survival of cancer patients diagnosed in Europe during 1990-94. Serial (Book, Monograph), 2003.
- [165] L. Ellis, B. Rachet, and M. P. Coleman. Cancer survival indicators for the national health service in england: the methodological implications of using cancer networks as geographic units of analysis. report for the national centre for health outcomes development. Report, 2006.
- [166] Glatte E. Atlas of cancer incidence in Norway 1970-1979. *Recent Results Cancer Res*, 114:216–226, 1989.

- [167] Pukkala E. Cancer maps of Finland: an example of small area-based mapping. *Recent Results Cancer Res*, 114:208–215, 1989.
- [168] Patama T and Pukkala E. Small-area based smoothing method for cancer risk mapping. *Spatial and Spatio-temporal Epidemiology*, 19:1–9, 2016. doi: <http://dx.doi.org/10.1016/j.sste.2016.05.003>.
- [169] Patama T, Engholm G, Larønningen S, Ólafsdóttir E, Khan S, Storm H, and Pukkala E. Small-area based map animations of cancer incidence in the nordic countries, 1971–2015, 2018. URL https://astra.cancer.fi/cancermaps/Nordic_18/.
- [170] Siesling S, van der Aa MA, Coebergh JW, Pukkala E, and Working Group of The Netherlands Cancer Registry. Time-space trends in cancer incidence in the Netherlands in 1989–2003. *International Journal of Cancer*, 122(9):2106–2114, May 2008.
- [171] Patama T. *Manual of Cancer Animation Mapping System*. Finnish Cancer Registry, 2011.
- [172] Egger M, Davey Smith G, Schneider M, and Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634, Sep 1997.
- [173] Sterne JA and Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54:1046–1055, 2001.
- [174] Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24:1185–1202, 2005.
- [175] Mayer EK, Bottle A, Rao C, Darzi AW, and Athanasiou T. Funnel plots and their emerging application in surgery. *Annals of Surgery*, 249:376–383, 2009.
- [176] D. J. Spiegelhalter. Handling over-dispersion of performance indicators. *Quality Safety Health Care*, 14:347–351, 2005.
- [177] Improving outcomes: A strategy for cancer. Technical report, Department of Health, 2012.
- [178] Pfeiffermann D. Small Area Estimation - new developments and directions. *International Statistical Review*, 70:125–143, 2002.

- [179] *Challenges in Statistics Production for Domains and Small Areas*, 2005. Small area international conference.
- [180] Rao JNK. *Small area estimation*. Wiley Series in Survey Methodology, 2003.
- [181] Särndal CE, Swensson B, and Wretman J. *Model Assisted Survey Sampling*. Springer Verlag, 1992.
- [182] Fay RE and Herriot RA. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal American Statistical Association*, 74:269–277, 1979.
- [183] Goldstein H, Browne W, and Rasbash J. Multilevel modelling of medical data. *Statistics in Medicine*, 21(21):3291–3315, 2002.
- [184] Goldstein H, Carpenter J, Kenward M, and Levin K. Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197, 2009.
- [185] Lawson AB. *Statistical Methods in Spatial Epidemiology*. Wiley, 2001.
- [186] Bell BS. Spatial analysis of disease-applications. *Cancer Treat Res*, 113:151–182, 2002.
- [187] Richardson S. *Spatial models in epidemiological applications. Highly Structured Stochastic Systems*. Oxford University Press, 2003.
- [188] Wakefield J and Elliott P. Issues in the statistical analysis of small area health data. *Statistics in Medicine*, 18(17-18):2377–2399, 1999.
- [189] Lawson AB and Williams FLR. *Introductory Guide to Disease Mapping*. Wiley, 2001.
- [190] Bithell JF. A classification of disease mapping methods. *Statistics in Medicine*, 19(17-18):2203–2215, 2000.
- [191] Knorr-Held L, Ralsser G, and Becker N. Disease mapping of stage-specific cancer incidence data. *Biometrics*, 58(3):492–501, 2002.
- [192] Marshall RJ. Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(2):283–294, 1991.

- [193] MacNab YC, Farrell PJ, Gustafson P, and Wen S. Estimation in Bayesian disease mapping. *Biometrics*, 60(4):865–873, 2004.
- [194] Buntinx F, Geys H, Lousbergh D, Broeders G, Cloes E, Dhollander D, Op De Beeck L, Vanden Brande J, Van Waes A, and Molenberghs G. Geographical differences in cancer incidence in the Belgian province of Limburg. *European Journal of Cancer*, 39(14):2058–2072, 2003.
- [195] Clayton DG, Bernardinelli L, and Montomoli C. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22:1193–1202, 1993.
- [196] Cressie N, Calder CA, Clark JS, Ver Hoef JM, and Wikle CK. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570, 2009.
- [197] Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.
- [198] Wakefield J. Ecologic studies revisited. *Annu Rev Public Health*, 29:75–90, 2008.
- [199] Knorr-Held L and Rasser G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.
- [200] Casseti T, La Rosa F, Rossi L, D’Alò D, and Stracci F. Cancer incidence in men: a cluster analysis of spatial patterns. *BMC Cancer*, 8:344, 2008.
- [201] Bilancia M and Fedespina A. Geographical clustering of lung cancer in the province of Lecce, Italy: 1992-2001. *International Journal of Health Geographics*, 8:40, 2009.
- [202] Kulldorff M and Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995.
- [203] Wakefield J and Kim A. A bayesian model for cluster detection. *Biostatistics*, 2013.
- [204] Cressie N and Read TRC. Spatial data analysis of regional counts. *Biometrical Journal*, 31:699–719, 1989.
- [205] Diggle P and Ribeiro PJR. *Model-based Geostatistics*. New York: Springer, 2007.

- [206] Goovaerts P. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics*, 4:31, 2005.
- [207] Goovaerts P. Geostatistical analysis of health data: state-of-the-art and perspective. *geoENV VI - Geostatistics for Environmental Applications*, pages 3–22, 2008.
- [208] Kafadar K. Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15(23):2539–2560, 1996.
- [209] Simonoff JS. *Smoothing Methods in Statistics*. Springer Verlag, New York, 1996.
- [210] Kafadar K. Simultaneous smoothing and adjusting mortality rates in U.S. counties: melanoma in white females and white males. *Statistics in Medicine*, 18(23):3167–3188, 1999.
- [211] Goovaerts P and Gebreab S. How does Poisson kriging compare to the popular BYM model for mapping disease risks? *International Journal of Health Geographics*, 7:6, 2008.
- [212] Berke O. Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(1):18, 2004.
- [213] Hampton KH, Serre ML, Gesink DC, Pilcher CD, and Miller WC. Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. *International Journal of Health Geographics*, 10:54, 2011.
- [214] Henderson R, Shimakura S, and Gorst D. Modeling Spatial Variation in Leukemia Survival Data. *Journal American Statistical Association*, 97:965–972, 2002.
- [215] Sauleau EA, Hennerfeind A, Buemi A, and Held L. Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geosadditive models. *Statistics in Medicine*, 26(1):212–229, 2007.
- [216] Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.
- [217] Bartlett MS. *An Introduction to Stochastic Processes*. Cambridge University Press, 1955.

- [218] Bartlett MS. A further note on nearest neighbour models. *Journal of the Royal Statistical Society, Series A*, 131:579–580, 1968.
- [219] Clayton D and Kaldor J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681, 1987.
- [220] Tsutakawa RK, Shoop GL, and Marienfeld CJ. Empirical bayes estimation of cancer mortality rates. *Statistics in Medicine*, 4(2):201–212, 1985.
- [221] Tsutakawa RK. Mixed model for analyzing geographic variability in mortality rates. *Journal American Statistical Association*, 83(401):37–42, 1988.
- [222] Carlin BP and Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman, 1996.
- [223] Leyland AH and Davies CA. Empirical bayes methods for disease mapping. *Statistical Methods Medical Research*, 14(1):17–34, 2005.
- [224] Besag J, York J, and Mollié A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43:1–59, 1991.
- [225] Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *Journal American Statistical Association*, 88:9–25, 1993.
- [226] Breslow N. Biostatistics and bayes. *Statistical Science*, 5(3):269–298, 1990.
- [227] Lee PM. *Bayesian Statistics: An Introduction*. Wiley, 2004.
- [228] Brooks SP. Markov chain Monte Carlo method and its application. *The Statistician*, 47:69–100, 1998.
- [229] Spiegelhalter D, Gilks WR, and Richardson S. *Markov chain Monte Carlo in practice*. Chapman, 1996.
- [230] Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, Schlattmann P, and Divino F. Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. *Statistics in Medicine*, 19(17-18):2217–2241, 2000.

- [231] Best N, Richardson S, and Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.
- [232] Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, and Songini M. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14 (21-22):2433–2443, 1995.
- [233] Knorr-Held L and Besag J. Modelling risk from a disease in time and space. *Statistics in Medicine*, 17(18):2045–2060, 1998.
- [234] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567, 2000.
- [235] MacNab YC and Dean CB. Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57(3):949–956, 2001.
- [236] MacNab YC and Dean CB. Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21(3):347–358, 2002.
- [237] Sun D, Tsutakawa RK, Kim H, and He Z. Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 19(15):2015–2035, 2000.
- [238] Tzala E and Best N. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, 17(1):97–118, 2008.
- [239] Langford IH, Marris C, McDonald AL, Goldstein H, Rasbash J, and O’Riordan T. Simultaneous analysis of individual and aggregate responses in psychometric data using multilevel modeling. *Risk Analysis*, 19(4):675–683, 1999.
- [240] Leyland AH, Langford IH, Rasbash J, and Goldstein H. Multivariate spatial models for event data. *Statistics in Medicine*, 19(17-18):2469–2478, 2000.
- [241] Assunção RM and Castro SM. Multiple cancer sites incidence rates estimation using a multivariate bayesian model. *International Journal of Epidemiology*, 33(3):508–516, 2004.
- [242] Knorr-Held L. Joint disease mapping. *In the Year of the Finnish Statistical Society*, pages 59–75, 1999-2000.

- [243] Downing A, Forman D, Gilthorpe MS, Edwards KL, and Manda SO. Joint disease mapping using six cancers in the Yorkshire region of England. *International Journal of Health Geographics*, 7:41, 2008.
- [244] Clayton DG. A Monte Carlo method for Bayesian inference in frailty. *Biometrics*, 47:467–485, 1991.
- [245] Banerjee S, Wall MM, and Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, 4(1):123–142, 2003.
- [246] S Banerjee, B.P. Carlin, and Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, 2004.
- [247] Vaupel JW, Manton KG, and Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.
- [248] Allemani C, Sant M, De Angelis R, Marcos-Gragera R, Coebergh JW, and EURO-CARE Working Group. Hodgkin disease survival in Europe and the US: prognostic significance of morphologic groups. *Cancer*, 107(2):352–360, 2006.
- [249] Engeland A, Haldorsen T, Dickman PW, Hakulinen T, Moller TR, Storm HH, and Tulinius H. Relative survival of cancer patients: a comparison between Denmark and other Nordic countries. *Acta Oncologica*, 37:49–59, 1998.
- [250] Osnes K and Aalen OO. Spatial smoothing of cancer survival: a Bayesian approach. *Statistics in Medicine*, 18:2087–2099, 1999.
- [251] Yu XQ, O'Connell DL, Giggerd RW, Smith DP, Dickman PW, and Armstron BK. Estimating regional variation in cancer survival: a tool for improving cancer care. *Cancer Causes and Control*, 15:611–618, 2004.
- [252] Cramb SM, Mengersen KL, and Baade PD. Developing the atlas of cancer in Queensland: methodological issues. *International Journal of Health Geographics*, 10: 9, 2011.

- [253] Fairley L, Forman D, West R, and Manda S. Spatial variation in prostate cancer survival in the Northern and Yorkshire region of England using Bayesian relative survival smoothing. *British Journal of Cancer*, 99(11):1786–1793, 2008.
- [254] Saez M, Barcelo MA, Martos C, Saurina C, Marcos-Gragera R, Renart G, Ocana-Riola R, Feja C, and Alcala T. Spatial variability in relative survival from female breast cancer. *Journal of the Royal Statistical Society, Series A*, 175:107–134, 2012.
- [255] Hennerfeind A, Held L, and Sauleau EA. A bayesian analysis of relative cancer survival with geoadditive models. *Statistical Modelling*, 8(2):117–139, 2008.
- [256] Cramb SM, Mengersen KL, Lambert PC, Ryan LM, and Baade PD. A flexible parametric approach to examining spatial variation in relative survival. *Statistics in Medicine*, 35(29):5448–5463, 2016.
- [257] Quaresma M, Carpenter J, and Rachet B. Flexible bayesian excess hazard models using low-rank thin plate splines. *Statistical Methods in Medical Research*, 2019. doi: 10.1177/0962280219874094.
- [258] NHS. Patient choice. URL <https://www.england.nhs.uk/patient-choice/>.
- [259] Crowther MJ and Lambert PC. A general framework for parametric survival analysis. *Statistics in Medicine*, 33:5280–5297, 2014.
- [260] Earnest A, Morgan G, Mengersen K, Ryan L, Summerhayes R, and Beard J. Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, 6:54, 2007.
- [261] Allemani C, Harewood R, Johnson CJ, Carreira H, Spika D, Bonaventure A, Ward K, Weir HK, and Coleman MP. Population based cancer survival in the United States: Data, quality control, and statistical methods. *Cancer*, 123(24):4982–4993, 2017.
- [262] Harrison CJ, Spencer RG, and Shackley DC. Transforming cancer outcomes in England: earlier and faster diagnoses, pathways to success, and empowering alliances. *Journal of Healthcare Leadership*, 11:1–11, 2019. doi: 10.2147/JHL.S150924.